Math Meth Oper Res (1999) 49:483–500

Mathematical Methods
of Operations Research

A decomposition approach for undiscounted two-person zero-sum stochastic games

Zeynep Müge Avşar[1], **Melike Baykal-Gürsoy**[2]

[1] Industrial Engineering Department, Bilkent University, Bilkent 06533, Ankara, Turkey (e-mail: avsar@bilkent.edu.tr)
[2] Industrial Engineering Department, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8018, USA (e-mail: gursoy@rci.rutgers.edu)

Abstract. Two-person zero-sum stochastic games are considered under the long-run average expected payoff criterion. State and action spaces are assumed finite. By making use of the concept of maximal communicating classes, the following decomposition algorithm is introduced for solving two-person zero-sum stochastic games: First, the state space is decomposed into maximal communicating classes. Then, these classes are organized in an hierarchical order where each level may contain more than one maximal communicating class. Best stationary strategies for the states in a maximal communicating class at a level are determined by using the best stationary strategies of the states in the previous levels that are accessible from that class. At the initial level, a restricted game is defined for each closed maximal communicating class and these restricted games are solved independently. It is shown that the proposed decomposition algorithm is exact in the sense that the solution obtained from the decomposition procedure gives the best stationary strategies for the original stochastic game.

Key words: Undiscounted stochastic games, decomposition

1 Introduction

In this article, two-person zero-sum stochastic games are considered. The players periodically observe the state of the process and independently take one of finitely many actions that are available at the current state. The state space is assumed finite. Depending on the current state and the actions taken, the state to be visited at the next epoch is determined and player II makes an instantaneous payment to player I. Under the long-run average expected

Manuscript received: June 1998/final version received: December 1998

payoff criterion, player II aims to minimize his average payment to player I who tries to maximize his average return. At each state, available actions of each player and the instantaneous payoff amounts and the transition probabilities that correspond to every action pair are all known by both of the players.

Stochastic games are classified according to their ergodic or data (e.g., transitions, payoffs) structure. The following are the stochastic games with different data structures: stochastic games with perfect information in which one player's action space is singleton at every state, the single-controller stochastic games, the switching-controller stochastic games, the separable reward state independent transition stochastic games (SER-SIT games), the additive reward additive transition stochastic games (AR-AT games). In this article, a game that is not unichain and does not have a specific data structure is going to be called as general stochastic game. Value of an undiscounted game is defined as the amount of long-run average expected payoff on which the players agree. A strategy pair that gives this payoff is called optimal strategy pair, none of the players can improve his payoff by unilateral deviations from this strategy pair. It is known that optimal strategies of the stochastic games are in the class of behavior strategies [1]. Stationary strategies form a subclass of behavior strategies. Depending on the current state of the process, stationary strategies are expressed by the probability distributions over the action spaces. It should be noted that a general stochastic game may not have an optimal stationary strategy pair. For such games, Filar et al. [9] defined best stationary strategies with respect to a measure of distance from optimality.

Focus of this article is on developing a decomposition procedure for general stochastic games. The aim is to compute best stationary strategies of any stochastic game by solving a number of stochastic games with smaller action and/or state spaces instead of solving the original game. For that purpose, what is proposed in this article is to decompose the state space and define a restricted game over each partition of the state space. It is shown that solutions of these restricted games give best stationary strategies for the original stochastic game.

State classification according to the accessibility relations was introduced by Bather [3]. As a result of this classification, Ross and Varadarajan [16] studied the concept of strongly communicating classes in detail. In [16], constrained Markov Decision Processes (MDPs) are solved by using a decomposition approach. Note that strongly communicating classes correspond to maximal recurrent classes. Later, Baykal-Gürsoy [4] employed the same concept for solving single-controller stochastic games. In these studies, the approach is to identify the strongly communicating classes and to decompose the state space into strongly communicating classes and a (possibly empty) set of transient states that are all disjoint. For each strongly communicating class, a system (a constrained MDP or a single-controller stochastic game) restricted to the states of that class, is defined. These restricted systems are solved independently. Then, an aggregate system is constructed based on the optimal or ε-optimal stationary strategies obtained from the restricted systems. Each strongly communicating class under its optimal stationary strategies is replaced with an aggregate state. Aggregate states together with the transient states form the aggregate system. Solution of this aggregate system gives optimal stationary strategies for the original constrained MDP or the original single-controller stochastic game.

The decomposition approach of [16] and [4] does not work for general stochastic games due to contradicting objectives of the players both of whom have control over the game. In this article, the following approach is introduced to solve undiscounted stochastic games: First, the state space is decomposed into maximal communicating classes, and then these classes are assigned to disjoint levels so that each maximal communicating class is considered at only one of those levels. The states are placed into levels in such a way that best stationary strategies for the states in maximal communicating classes considered at a level are determined by the best stationary strategies of the states considered in the previous levels. A restricted game is constructed for each maximal communicating class over the state space of that class as well as the states at the previous levels that are accessible from that class. Recurrent classes that are formed under the best stationary strategies for the restricted games of the previous levels are replaced with aggregate states while keeping the transient states as they are. Starting from the initial level of the hierarchy, each restricted game is solved independently. Thus, the restricted games solved at a level give best stationary strategies for the states of that level. In solving restricted games, the ergodic and/or data structure of the game could be exploited and efficient algorithms could be used (an extensive survey of the existing algorithms is given in [13]).

This article is organized as follows: In section 2, notation is introduced. Section 3 introduces the proposed decomposition approach. In section 4, construction of the restricted games is explained and the proposed algorithm is given. It is shown that the stationary strategies obtained by the decomposition algorithm are the best stationary strategies of the original stochastic game.

2 Preliminaries

The underlying stochastic process for a two-person zero-sum stochastic game is $\{(X_n, A_n, B_n), n = 1, 2, \ldots\}$ where X_n and A_n (B_n) are the random variables that denote the state of the game and the action taken by player I (II), respectively, at decision epoch n. X_n takes values in a finite state space $\mathscr{S} = \{1, \ldots, S\}$, say $X_n = i$. At state i, player I (II) takes an action, say a (b), from a finite action space $\mathscr{A}_i = \{1, \ldots, M_i\}$ ($\mathscr{B}_i = \{1, \ldots, N_i\}$). The amount of instantaneous payment made by player II to player I at epoch n is denoted by the random variable $R_n = R(X_n, A_n, B_n)$ as a function of the state visited and the actions taken at this epoch. Payoff amounts are finite. The next state of the process is determined via the transition probabilities that are also called the law of motion. The process is assumed time-homogeneous, i.e., the expected payoff $E(R(X_n = i, A_n = a, B_n = b))$ is equal to r_{iab} and the transition probability $P(X_{n+1} = j \mid X_n = i, A_n = a, B_n = b)$ is equal to P_{iabj} for every n. Stationary strategies of players I and II are denoted by the vectors $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1M_1}; \alpha_{21}, \ldots, \alpha_{2M_2}; \ldots; \alpha_{S1}, \ldots, \alpha_{SM_S})$ and $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{1N_1}; \beta_{21}, \ldots, \beta_{2N_2}; \ldots; \beta_{S1}, \ldots, \beta_{SN_S})$, respectively. Note that α_{ia} (β_{ib}) is the conditional probability $P(A_n = a \mid X_n = i)$ ($P(B_n = b \mid X_n = i)$). The stationary strategy pair taken by the players in state i, i.e., $((\alpha_{i1}, \ldots, \alpha_{iM_i}), (\beta_i, \ldots, \beta_{iN_i}))$, is denoted as $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ for every $i \in \mathscr{S}$. If stationary strategies $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are assigned to the players, the expected payoffs and the transition probabilities under these strategies are represented by $r_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $P_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, respectively.

When the initial state of the process is i, the long-run average expected payoff is denoted by $\phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as a function of the stationary strategy pair $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ taken by the players and it is defined as follows:

$$\phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} E_{\alpha,\beta}(R_n \mid X_1 = i),$$

where $E_{\alpha,\beta}$ represents the expectation under stationary strategy pair $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

If there exists a stationary strategy pair $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ that satisfies the saddle point condition, i.e., $\phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}^*) \leq \phi_i(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \leq \phi_i(\boldsymbol{\alpha}^*, \boldsymbol{\beta})$ for every $i \in \mathscr{S}$ and all stationary strategies $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, then $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is called optimal. As the optimality condition implies, unilateral deviations of player I (II) from $\boldsymbol{\alpha}^*$ ($\boldsymbol{\beta}^*$) results in less reward (more loss) for him. The corresponding payoff $\phi_i(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is called the value of the game for initial state i. As an immediate implication of the saddle point condition, ε-optimality is defined as follows: A stationary strategy pair $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is said to be ε-optimal if $\phi_i(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}}) - \varepsilon \leq \phi_i(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) \leq \phi_i(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}) + \varepsilon$ holds true for every $i \in \mathscr{S}$ and all stationary strategies $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Best stationary strategies are defined via the use of a distance function introduced by Filar et. al [9]. This distance function δ can be evaluated for any stationary strategy pair $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ by making use of the following formula: $\delta(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) = \sum_{i \in \mathscr{S}} (\max_{\alpha} \phi_i(\boldsymbol{\alpha}, \tilde{\boldsymbol{\beta}}) - \min_{\beta} \phi_i(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}))$. Clearly, δ is always nonnegative since each summation term is nonnegative. $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is said to be ε-optimal if $\delta(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is less than or equal to ε. Hence, $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is optimal if δ vanishes at this point.

At each state i, the parameters of the game are given by matrix $\begin{bmatrix} r_{iab} & \\ & j \end{bmatrix}$ where rows (columns) correspond to the actions available for player I (II). j is the state to be visited at the next epoch given that the current state is i and players I and II take actions a and b, respectively. If the next state is determined according to a probability distribution over the state space, then this distribution is written in the lower right corner.

3 Decomposition in stochastic games

This section starts with the definitions of maximal and strongly communicating classes in stochastic games and an adaptation of a procedure in [16] to identify strongly communicating classes.

A communicating class is called maximal communicating class if it is the largest obtainable under every possible stationary strategy pair. The maximal communicating classes could be open, i.e., a maximal communicating class may have transitions to states out of the class, or closed. Clearly, maximal communicating classes that are closed are maximal recurrent classes. If a state visited by the process is left in one transition with probability 1 under every stationary strategy pair, it defines a maximal communicating class by itself. Let $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_\Omega$ be the maximal communicating classes. The collection of maximal communicating classes $\{\mathscr{D}_1, \ldots, \mathscr{D}_\Omega\}$ defines a (unique) partition of the state space.

Definition 1. *A set of states $\mathscr{C} \subseteq \mathscr{S}$ is called a strongly communicating class if*

(i) $\mathscr{C}$ is recurrent under some stationary strategy pair, (ii) $\mathscr{C}$ is not a proper subset of another set that satisfies (i).

Obviously, strongly communicating classes are maximal recurrent classes. Note that maximal recurrent classes may not be recurrent under every stationary strategy pair. Let $\mathscr{C}_1, \ldots, \mathscr{C}_K$ denote the strongly communicating classes. The states that are not in strongly communicating classes are transient under every stationary strategy pair. Let $\mathscr{H}$ be the set of these transient states. By definition, $\mathscr{H} = \mathscr{S} - (\bigcup_{k=1}^{K} \mathscr{C}_k)$. The following lemma is analogous to an observation for MDPs due to Ross and Varadarajan [16].

Lemma 1. *The collection of strongly communicating classes and the set of transient states $\{\mathscr{C}_1, \ldots, \mathscr{C}_K, \mathscr{H}\}$ forms a (unique) partition of the state space $\mathscr{S}$.*

Proof: The collection $\{\mathscr{C}_1, \ldots, \mathscr{C}_K, \mathscr{H}\}$ covers $\mathscr{S}$, and the set of strongly communicating classes and $\mathscr{H}$ are disjoint by definition. So, it has to be shown that the strongly communicating classes are disjoint. Suppose there exist two strongly communicating classes, $\mathscr{C}_1$ and $\mathscr{C}_2$, that are not disjoint. Using some stationary strategy pairs $(\boldsymbol{\alpha}(1), \boldsymbol{\beta}(1))$ and $(\boldsymbol{\alpha}(2), \boldsymbol{\beta}(2))$ under which $\mathscr{C}_1$ and $\mathscr{C}_2$ are recurrent, respectively, let

$$\tilde{\alpha}_{ia} = \begin{cases} \alpha_{ia}(1) & i \in (\mathscr{C}_1 - \mathscr{C}_2), a \in \mathscr{A}_i, \\ \alpha_{ia}(2) & i \in (\mathscr{C}_2 - \mathscr{C}_1), a \in \mathscr{A}_i, \\ \lambda\alpha_{ia}(1) + (1-\lambda)\alpha_{ia}(2) & i \in (\mathscr{C}_1 \cap \mathscr{C}_2), a \in \mathscr{A}_i, \end{cases}$$

where $0 < \lambda < 1$. Define $\tilde{\boldsymbol{\beta}}$ similarly. Now, consider the state set $\mathscr{C} = \mathscr{C}_1 \cup \mathscr{C}_2$. Since the states in $\mathscr{C}$ are accessible from each other in $P(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$, $\mathscr{C}$ is said to be communicating under $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$. Also, both $\mathscr{C}_1$ and $\mathscr{C}_2$ are recurrent under $(\boldsymbol{\alpha}(1), \boldsymbol{\beta}(1))$ and $(\boldsymbol{\alpha}(2), \boldsymbol{\beta}(2))$, respectively, which leads to $\sum_{j \in (\mathscr{S} - \mathscr{C})} P_{ij}(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}) = 0$, $i \in \mathscr{C}$. Then, $\mathscr{C}$ is recurrent under $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$. Thus, $\mathscr{C}_1$ and $\mathscr{C}_2$ are proper subsets of $\mathscr{C}$ that satisfies (i) in definition 1. This contradicts the assumption that $\mathscr{C}_1$ and $\mathscr{C}_2$ are strongly communicating classes. □

In [16], Ross and Varadarajan outlined a procedure to decompose state space of an MDP into strongly communicating classes and transient states. Here, this procedure is revised for stochastic games. The original stochastic game is considered at the first step. The maximal communicating classes of the state space are identified. If a maximal communicating class is closed, then this set is labeled as a strongly communicating class. Since the state space is assumed finite, there exists at least one closed communicating class. If there are open communicating classes, these classes are considered one by one. Suppose $\mathscr{D}$ is such a class. If there exists an action pair (a, b) that takes the process out of class $\mathscr{D}$ from a state of class $\mathscr{D}$, say state i, i.e., $\sum_{j \in (\mathscr{S} - \mathscr{D})} P_{iabj} > 0$, then the corresponding entry in the matrix of state i is deleted. Note that not necessarily the actions a and b but the transitions expressed by distribution $(P_{iab1}, \ldots, P_{iabS})$ are deleted. If state i is left with an empty action space, then it is labeled as a transient state. This procedure is repeated for every state in $\mathscr{D}$

until a set $\mathscr{D}' \subseteq \mathscr{D}$ is obtained with the following properties: (i) there is at least one action pair for every state of $\mathscr{D}'$; (ii) none of the states in $(\mathscr{S} - \mathscr{D}')$ is reachable from $\mathscr{D}'$ under the remaining action pairs. Note that $\mathscr{D}'$ is obtained when the transitions specified above are deleted. At the second step, the same procedure is employed for $\mathscr{D}'$ with the remaining transitions of the states in $\mathscr{D}'$, i.e., each closed maximal communicating class is labeled as a strongly communicating class and each open maximal communicating class is examined separately. This is repeated until every state in $\mathscr{S}$ is either labeled as a transient state or considered in a strongly communicating class. The set of transient states is further decomposed into maximal communicating classes.

In the example problem given below, maximal and strongly communicating classes are identified.

Example 1

a) Consider the following stochastic game with nine states.

$$\begin{bmatrix} 5 & 9 \\ 2 & 1 \\ 7 & 3 \\ 1 & 1 \end{bmatrix} \quad i = 1 \qquad \begin{bmatrix} 4 \\ 2 \\ 10 \\ 1 \end{bmatrix} \quad i = 2 \qquad \begin{bmatrix} 2 \\ 3 \\ 1 \\ 3 \end{bmatrix} \quad i = 3$$

$$\begin{bmatrix} 20 \\ (0, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, 0, 0, 0) \end{bmatrix} \quad i = 4 \qquad \begin{bmatrix} 1 & 9 \\ 5 & 6 \\ 14 & 21 \\ 2 & 3 \end{bmatrix} \quad i = 5 \qquad \begin{bmatrix} 11 \\ 5 \end{bmatrix} \quad i = 6$$

$$\begin{bmatrix} 8 & 1 \\ 4 & 2 \end{bmatrix} \quad i = 7 \qquad \begin{bmatrix} 10 & 3 \\ 6 & 8 \\ 6 & 5 \\ 8 & 3 \end{bmatrix} \quad i = 8 \qquad \begin{bmatrix} 1 & 3 \\ 4 & 8 \\ 8 & 2 \\ 9 & 4 \end{bmatrix} \quad i = 9$$

The maximal communicating classes are $\mathscr{D}_1 = \{1, 2\}$, $\mathscr{D}_2 = \{3\}$, $\mathscr{D}_3 = \{4\}$, $\mathscr{D}_4 = \{5, 6\}$, $\mathscr{D}_5 = \{7\}$, $\mathscr{D}_6 = \{8\}$, $\mathscr{D}_7 = \{9\}$. $\mathscr{D}_1$ and $\mathscr{D}_2$ are closed whereas $\mathscr{D}_3$, $\mathscr{D}_4$, $\mathscr{D}_6$ and $\mathscr{D}_7$ are open. The strongly communicating classes are $\mathscr{C}_1 = \{1, 2\}$, $\mathscr{C}_2 = \{3\}$, $\mathscr{C}_3 = \{5, 6\}$, $\mathscr{C}_4 = \{8\}$, $\mathscr{C}_5 = \{9\}$ and the set of transient states is $\mathscr{H} = \{4, 7\}$. Note that in this example every strongly communicating class is also a maximal communicating class. □

b) Consider the stochastic game for which the matrices of states 6, 7 and 8 are given as

$$\begin{bmatrix} 11 \\ (0,0,\frac{2}{5},0,\frac{3}{5},0,0,0,0) \end{bmatrix} \quad \begin{bmatrix} 8 & 1 \\ 4 & (0,\frac{1}{4},0,0,0,0,\frac{1}{4},\frac{1}{2},0) \end{bmatrix}$$

$i = 6$ $\qquad$ $i = 7$

$$\begin{bmatrix} 10 & & & 3 \\ 6 & & & 7 \\ 6 & & & 5 \\ (0,0,0,0,0,0,\frac{1}{3},\frac{2}{3},0) & & & 3 \end{bmatrix}$$

$i = 8$

and the matrices for the remaining states stay the same as in part (a). Then, the maximal communicating classes are $\mathscr{D} = \{1,2\}$, $\mathscr{D}_2 = \{3\}$, $\mathscr{D}_3 = \{4\}$, $\mathscr{D}_4 = \{5,6\}$, $\mathscr{D}_5 = \{7,8\}$, $\mathscr{D}_6 = \{9\}$. $\mathscr{D}_1$ and $\mathscr{D}_2$ are closed unlike $\mathscr{D}_3$, $\mathscr{D}_4$, $\mathscr{D}_5$, $\mathscr{D}_6$. The strongly communicating classes are $\mathscr{C}_1 = \{1,2\}$, $\mathscr{C}_2 = \{3\}$, $\mathscr{C}_3 = \{5\}$, $\mathscr{C}_4 = \{9\}$ and the set of transient states is $\mathscr{H} = \{4,6,7,8\}$. □

Next, an example is given to demonstrate why the solution approach proposed in [4] for single-controller stochastic games does not work for stochastic games in general. This is a special case of a stochastic game with switching controllers, where at every state i either $P_{iabj} = P_{iaj}$, meaning only player I controls the law of motion, or $P_{iabj} = P_{ibj}$.

Example 2

$$\begin{bmatrix} 10 & 10 & 14 \\ 2 & 3 & 4 \end{bmatrix} \quad \begin{bmatrix} 7 & \\ & 2 \\ 6 & \\ & 1 \\ 10 & \\ & 4 \end{bmatrix} \quad \begin{bmatrix} 2 & \\ & 1 \\ 10 & \\ & 3 \\ 5 & \\ & 5 \end{bmatrix} \quad \begin{bmatrix} 16 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 15 \\ 5 \end{bmatrix}$$

$i = 1$ $\qquad$ $i = 2$ $\qquad$ $i = 3$ $\qquad$ $i = 4$ $\qquad$ $i = 5$

Open and closed strongly communicating classes are $\{1,2,3\}$ and $\{4\}$, $\{5\}$, respectively. The value of the game that is restricted to the state space $\{4\}$ ($\{5\}$) is equal to 16 (15). The game restricted to the state space $\{1,2,3\}$ is

$$\begin{bmatrix} 10 & 10 \\ 2 & 3 \end{bmatrix} \quad \begin{bmatrix} 7 & \\ & 2 \\ 6 & \\ & 1 \end{bmatrix} \quad \begin{bmatrix} 2 & \\ & 1 \\ 10 & \\ & 3 \end{bmatrix}.$$

$i = 1$ $\qquad$ $i = 2$ $\qquad$ $i = 3$

The value of this restricted game is 10 (8) for initial state 3 (1 or 2), and the optimal stationary strategies are $\boldsymbol{\alpha}^* = (1; 0, 1; 0, 1)$ and $\boldsymbol{\beta}^* = (1, 0; 1; 1)$. In [4], each strongly communicating class is replaced with an aggregate state because the value of a restricted game defined over a strongly communicating class is independent of the initial state. In this example problem, the value of the restricted game defined over the state space $\{1, 2, 3\}$ depends on the initial state. Hence, the decomposition procedure in [4] can not be applied directly but one might think of employing the idea by replacing each subset of $\{1, 2, 3\}$ that is recurrent under the optimal stationary strategies of the restricted game with an aggregate state, and then constructing an aggregate game considering the transitions that take the process out of the aggregate states. For this example problem, $\{1, 2\}$ and $\{3\}$ are recurrent under the optimal stationary strategies of the restricted game; so, each can be replaced with an aggregate state by defining absorbing action pairs with the corresponding payoff amounts 8 and 10, respectively. Then, the next step would be to construct the aggregate game with the use of transitions $(1, 2)$ and $(1, 3)$ of state 1 and $(3, 1)$ of state 2 ((1,1) and (3,1) of state 3) which can take the process out of the aggregate state $\{1, 2\}$ ($\{3\}$). One problem related with this approach is to construct such an aggregate game; this is not an easy task as in the case of constrained MDPs and the single-controller stochastic games as observed with the use of this example. Even if this difficulty can be handled, the solution of the aggregate game would not give the best stationary strategies for the original game because the value of the original game is 16 (15) for initial state 2 (1 or 3), i.e., the original game value of state 1 is different from the value of state 2 although these two states are considered to define an aggregate state in the aggregate game.

One other extension of the idea in [4] might be to fix optimal stationary strategies of the restricted games. As another example, consider the case where $r_{221} = 3$. Then, the value of the restricted game defined over $\{1, 2, 3\}$ is 10 (7) for initial state 3 (1 or 2) and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = ((1; 1, 0; 0, 1), (1, 0; 1; 1))$. Consider the overall game by fixing optimal strategies of the strongly communicating classes:

$$\underset{i=1}{\begin{bmatrix} 10 & 14 \\ 2 & 4 \end{bmatrix}} \quad \underset{i=2}{\begin{bmatrix} 7 \\ 2 \\ 10 \\ 4 \end{bmatrix}} \quad \underset{i=3}{\begin{bmatrix} 10 \\ 3 \\ 5 \\ 5 \end{bmatrix}} \quad \underset{i=4}{\begin{bmatrix} 16 \\ 4 \end{bmatrix}} \quad \underset{i=5}{\begin{bmatrix} 15 \\ 5 \end{bmatrix}}.$$

The solution of this game gives a value of 15 (16) for the initial state 3 (1 or 2). When the process is initially in 4 (5), the game value stays as 16 (15) because $\{4\}$ ($\{5\}$) is a closed class. However, the value of the original stochastic game is 16 (15) for the initial state 2 (1 or 3) under optimal stationary strategies $\boldsymbol{\alpha}^* = (1; 0, 0, 1; 0, 0, 1; 1; 1)$ and $\boldsymbol{\beta}^* = (0, 1, 0; 1; 1; 1; 1)$. Thus, this approach does not give the correct result when the initial state is 1. □

Note that, unlike constrained MDPs and single-controller stochastic games studied in [15], [16] and [4], respectively, under the best stationary strategy pair a strongly communicating class (even a closed communicating stochastic

game) may have a multichain structure with more than one recurrent class having different average payoff amounts and (maybe) some transient states. An example of this case for a (closed) communicating stochastic game is given in [2].

From the analysis of example 2, it is observed that the solution of each game restricted to a strongly communicating class (even to a maximal communicating class) may not lead to the best stationary strategies for the initial states in that class. This is because both of the players have control over the game unlike single-controller stochastic games. In this article, a new solution procedure is introduced to solve general stochastic games. First, the state space is decomposed into maximal communicating classes. Further, these maximal communicating classes are decomposed into hierarchically ordered disjoint levels. The levels of the classes are determined according to the following definition:

Definition 2. *If a maximal communicating class is closed, then its level is* 0. *If a maximal communicating class is open, its level n is the maximum number of transitions it takes to reach a level* 0 *class without counting more than one visit to a maximal communicating class.*

Let L_n denote the set of states in the maximal communicating classes at level n. The levels in example 1(a) are $L_0 = \mathscr{D}_1 \cup \mathscr{D}_2$, $L_1 = \mathscr{D}_3 \cup \mathscr{D}_4$, $L_2 = \mathscr{D}_5 \cup \mathscr{D}_6$, $L_3 = \mathscr{D}_7$. In example 1(b), the levels are $L_0 = \mathscr{D}_1 \cup \mathscr{D}_2$, $L_1 = \mathscr{D}_3 \cup \mathscr{D}_4$, $L_2 = \mathscr{D}_5$, $L_3 = \mathscr{D}_6$. Note that L_n is a disjoint set, i.e., there does not exist any action pair under which a maximal communicating class is accessible from another class at the same level. This property allows one to solve an independent game restricted to a maximal communicating class at a level and all other states at the previous levels that are accessible from this class.

Next, a procedure is given to identify the levels of a stochastic game.

Step 1) Let L_0 be the union of closed maximal communicating classes, and let $n = 0$. If $L_0 = \mathscr{S}$, stop. Otherwise, go to step 2.
Step 2) Identify each maximal communicating class $\mathscr{D}$ in $\left(\mathscr{S} - \left(\bigcup_{d=0}^{n} L_d\right)\right)$ such that $\sum_{j \in L_n} P_{iabj} > 0$ for some i in $\mathscr{D}$ and $(a, b) \in \mathscr{A}_i \times \mathscr{B}_i$. Let $\mathscr{G}$ be the set of states in such classes.
Step 3) For a maximal communicating class $\mathscr{D} \subseteq \mathscr{G}$, if $\sum_{j \in (\mathscr{G} - \mathscr{D})} P_{iabj} = 0$ for every $i \in \mathscr{D}$, $(a, b) \in \mathscr{A}_i \times \mathscr{B}_i$, then put $\mathscr{D}$ in L_{n+1}.
Step 4) If $\left(\bigcup_{d=0}^{n+1} L_d\right) = \mathscr{S}$, stop. Otherwise, increment n by 1 *and go to step 2.*

4 The proposed procedure

The algorithm proposed in this study can be outlined as follows: At level 0, the stochastic games restricted to the closed maximal communicating classes are solved independently to obtain the best stationary strategies of the states in these classes. Based on their ergodic structure under the best stationary strategies, recurrent classes formed are replaced with absorbing aggregate states. At the next level, for each class in L_1 a restricted game is constructed by fixing the best stationary strategies of the states in L_0. The state space of the restricted game is composed of the states in that class together with the

aggregate and transient states of level 0 that are accessible from these states. Solutions of these restricted games give the best stationary strategies for the states in L_1. Then, using these solutions the algorithm proceeds to the next level until every class in $\mathscr{S}$ is taken into consideration.

Under any stationary strategy pair, each recurrent class is in one of the strongly communicating classes (the arguments that lead to this result are presented in [16]) which are subsets of the maximal communicating classes. But all of the states in a maximal communicating class may be transient under every stationary strategy pair. Considering these and the communication property satisfied by the maximal communicating classes, in the proposed decomposition algorithm each maximal communicating class is considered at one of the disjoint levels. Then, every recurrent class obtained under the best stationary strategies of a level can be replaced with an absorbing aggregate state in the restricted games of the next levels.

In this study, there is no condition imposed on the ergodic properties and/or the data (i.e., transitions and/or payoffs) of the stochastic games. When the original or a restricted game falls into a class of stochastic games with specific ergodic and/or data structure, an algorithm that exploits the special structure may be used in the implementation of the decomposition procedure. Various algorithms are available in the literature for irreducible stochastic games (Hoffman and Karp [10]), unichain stochastic games (Federgruen [5], Van der Wal [17]), stochastic games with a value independent of the initial state (Federgruen [5]), communicating stochastic games with a value independent of the initial state (Avşar and Baykal-Gürsoy [2]), stochastic games with perfect information and single-controller (Filar [6], Vrieze [18], Hordijk and Kallenberg [11], Baykal-Gürsoy [4]), switching-controller stochastic games (Filar and Raghavan [7], Vrieze et. al [19]), SER-SIT games (Parthasarathy et. al. [12]), AR-AT games (Raghavan and Tijs and Vrieze [14]). If a given stochastic game does not fall into any of these classes, then the NLP formulation due to Filar et. al. [9] can be used. In such a case, the proposed decomposition algorithm would make the use of this NLP easier, especially for the stochastic games with large state and/or action spaces. Since the NLP formulation works for every stochastic game regardless of its ergodic and data structure, the proposed approach will be explained via the use of this NLP formulation.

NLP formulation in [9] is based on a characterization of the stationary equilibrium due to Filar and Schultz [8] and it finds the best stationary strategies even when the optimal stationary strategies fail to exist. This formulation is given below. It supplies the best stationary strategy pair with respect to the measure δ.

Problem 1

$$\inf \sum_{i \in \mathscr{S}} (g_i - u_i)$$

$$s.t. \quad g_i \geq \sum_{j \in \mathscr{S}} P_{iaj}(\boldsymbol{\beta}) g_j, \quad i \in \mathscr{S}, a \in \mathscr{A}_i,$$

$$g_i + v_i \geq r_{ia}(\boldsymbol{\beta}) + \sum_{j \in \mathscr{S}} P_{iaj}(\boldsymbol{\beta}) v_j, \quad i \in \mathscr{S}, a \in \mathscr{A}_i,$$

$$u_i \leq \sum_{j \in \mathscr{S}} P_{ibj}(\boldsymbol{a}) u_j, \quad i \in \mathscr{S}, b \in \mathscr{B}_i,$$

$$u_i + t_i \leq r_{ib}(\boldsymbol{a}) + \sum_{j \in \mathscr{S}} P_{ibj}(\boldsymbol{a}) t_j, \quad i \in \mathscr{S}, b \in \mathscr{B}_i,$$

$$\sum_{a \in \mathscr{A}_i} \alpha_{ia} = 1, \quad i \in \mathscr{S},$$

$$\alpha_{ia} \geq 0, \quad i \in \mathscr{S}, a \in \mathscr{A}_i,$$

$$\sum_{b \in \mathscr{B}_i} \beta_{ib} = 1, \quad i \in \mathscr{S},$$

$$\beta_{ib} \geq 0, \quad i \in \mathscr{S}, b \in \mathscr{B}_i,$$

$$g_i, u_i, v_i, t_i \textit{ unrestricted}, \quad i \in \mathscr{S},$$

where the decision variables $g_i(u_i)$ and $v_i(t_i)$ are the long-run average expected payoff and the change in the total payoff, respectively, when the second (first) player employs stationary strategy $\boldsymbol{\beta}(\boldsymbol{a})$ and the initial state is $i, i \in \mathscr{S}$. Note that the objective function gives the value of the distance function δ at $(\boldsymbol{a}, \boldsymbol{\beta})$. If the optimal objective function value is zero, then the stochastic game has optimal stationary strategies. On the other hand, for an ε-optimal stationary strategy pair an upper bound on the objective function value is $2S\varepsilon$. Another observation that results from *Problem 1* is that if the minimum of the objective function does not exist, then for every $\varepsilon > 0$ there exists a stationary strategy pair that is $(\varepsilon + \inf \sum_{i \in \mathscr{S}} (g_i - u_i))$-optimal.

Remark: Problem 1 is separable. Minimization of $\sum_{i \in \mathscr{S}} g_i$ over the constraints in terms of $\boldsymbol{g}, \boldsymbol{v}, \boldsymbol{\beta}$ and maximization of $\sum_{i \in \mathscr{S}} u_i$ over the constraints in terms of $\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{a}$ are two bilinear problems that are independent of each other. The former subproblem maximizes payoff over $\boldsymbol{a}$ strategies for a given $\boldsymbol{\beta}$ and minimizes this amount over all $\boldsymbol{\beta}$ strategies. So, it solves $\min_\beta \max_\alpha \phi_i(\boldsymbol{a}, \boldsymbol{\beta})$, $i \in \mathscr{S}$, for the best $\boldsymbol{\beta}$ strategy. Similarly, the latter subproblem solves $\max_\alpha \min_\beta \phi_i(\boldsymbol{a}, \boldsymbol{\beta})$, $i \in \mathscr{S}$, for the best $\boldsymbol{a}$ strategy. For the former subproblem, let $\boldsymbol{\beta}^*$ be the solution and $\hat{\boldsymbol{a}}^*$ be the maximizing strategy of the first player given $\boldsymbol{\beta}^*$ as the second player's strategy. Then, $(\hat{\boldsymbol{a}}^*, \boldsymbol{\beta}^*)$ satisfies $\min_\beta \max_\alpha \phi_i(\boldsymbol{a}, \boldsymbol{\beta}) = \phi_i(\hat{\boldsymbol{a}}^*, \boldsymbol{\beta}^*)$, $i \in \mathscr{S}$. Similarly, let $(\boldsymbol{a}^*, \hat{\boldsymbol{\beta}}^*)$ be the solution for the latter subproblem. Since the existence of optimal stationary strategies is not presumed in this article, this property, i.e., characterization of the best stationary strategies by two independent programs, shows the requirement to use the decomposition procedure two times. One pass is needed to compute $\boldsymbol{g}$ and $\boldsymbol{\beta}$ vectors and the other pass is to find $\boldsymbol{u}$ and $\boldsymbol{a}$ vectors. In the former (latter) pass, the restricted games are solved for $\min_\beta \max_\alpha \phi_i(\boldsymbol{a}, \boldsymbol{\beta})$ $(\max_\alpha \min_\beta \phi_i(\boldsymbol{a}, \boldsymbol{\beta}))$ values at every level.

If the process is initially at level n, *Problem 1* is reduced to the following formulation in order to find the best stationary strategies corresponding to the states in L_n:

Problem 2_n

$$\text{Min} \sum_{i \in \left(\bigcup_{d=0}^{n} L_d\right)} (g_i - u_i)$$

$$s.t.\quad g_i \geq \sum_{j \in \left(\bigcup_{w=0}^{d} L_w\right)} P_{iaj}(\boldsymbol{\beta}) g_j, \quad i \in L_d, a \in \mathscr{A}_i, d = 0, \ldots, n,$$

$$g_i + v_i \geq r_{ia}(\boldsymbol{\beta}) + \sum_{j \in \left(\bigcup_{w=0}^{d} L_w\right)} P_{iaj}(\boldsymbol{\beta}) v_j, \quad i \in L_d, a \in \mathscr{A}_i, d = 0, \ldots, n,$$

$$u_i \leq \sum_{j \in \left(\bigcup_{w=0}^{d} L_w\right)} P_{ibj}(\boldsymbol{\alpha}) u_j, \quad i \in L_d, b \in \mathscr{B}_i, d = 0, \ldots, n,$$

$$u_i + t_i \leq r_{ib}(\boldsymbol{\alpha}) + \sum_{j \in \left(\bigcup_{w=0}^{d} L_w\right)} P_{ibj}(\boldsymbol{\alpha}) t_j, \quad i \in L_d, b \in \mathscr{B}_i, d = 0, \ldots, n,$$

$$\sum_{a \in \mathscr{A}_i} \alpha_{ia} = 1, \quad i \in \left(\bigcup_{d=0}^{n} L_d\right),$$

$$\alpha_{ia} \geq 0, \quad i \in \left(\bigcup_{d=0}^{n} L_d\right), a \in \mathscr{A}_i,$$

$$\sum_{b \in \mathscr{B}_i} \beta_{ib} = 1, \quad i \in \left(\bigcup_{d=0}^{n} L_d\right),$$

$$\beta_{ib} \geq 0, \quad i \in \left(\bigcup_{d=0}^{n} L_d\right), b \in \mathscr{B}_i,$$

$$g_i, u_i, v_i, t_i \textit{ unrestricted}, \quad i \in \left(\bigcup_{d=0}^{n} L_d\right).$$

Without making use of the aggregation concept, for each level such reductions in *Problem 1* show why the proposed decomposition procedure gives the best stationary strategies of a stochastic game. This observation is stated in proposition 1.

Proposition 1 *Problem* 2_n *gives best stationary strategies for the states in* $\left(\bigcup_{d=0}^{n} L_d\right)$.

Proof: From definition 2, for every $i \in L_n$

$$\sum_{j \in \mathscr{S}} P_{iabj} = \sum_{j \in \left(\bigcup_{d=0}^{n} L_d\right)} P_{iabj} = 1, \quad (a, b) \in \mathscr{A}_i \times \mathscr{B}_i,$$

which means that for computing best stationary strategies of the states at level n it is sufficient to consider the stochastic game defined over the state space $\left(\bigcup_{d=0}^{n} L_d\right)$. Hence, *Problem* 2_n gives $(\boldsymbol{\alpha}_i^*, \boldsymbol{\beta}_i^*)$ and $\hat{\boldsymbol{\alpha}}_i^*, \hat{\boldsymbol{\beta}}_i^*$ for every $i \in \left(\bigcup_{d=0}^{n} L_d\right)$. □

Note that each maximal communicating class is kept as a partition in the

decomposed state space because under every stationary strategy pair the recurrent classes are subsets of the maximal communicating classes. Then, since L_n is a disjoint set, each maximal communicating class in L_n and the states that are accessible from it can be considered in an independent problem with all the variables and the constraints in *Problem* 2_n corresponding to these states. This observation leads to the introduction of a restricted game for each maximal communicating class in L_n with all the states that can be reached from this class. In the following subsection, the construction of a restricted game at a level is explained. This construction is based on the ergodic structure of the game that is determined by the best stationary strategies of the states at the previous levels. Next, the proposed algorithm is presented and it is shown that this algorithm finds the best stationary strategies of the stochastic games.

4.1 The restricted games

For every closed (absorbing) maximal communicating class $\mathscr{D}_m$ at level 0, a restricted game is defined over state space $\mathscr{D}_m$ with action spaces $\mathscr{A}_i$ and $\mathscr{B}_i$ for $i \in \mathscr{D}_m$. Let $(\boldsymbol{\alpha}^m, \boldsymbol{\beta}^m)$ be the best stationary strategy pair and $\hat{\boldsymbol{\alpha}}^m(\hat{\boldsymbol{\beta}}^m)$ be the strategy maximizing (minimizing) $\phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $i \in \mathscr{S}$, given $\boldsymbol{\beta}^m(\boldsymbol{\alpha}^m)$ for the second (first) player. Strategy pairs $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ and $(\boldsymbol{\alpha}^m, \hat{\boldsymbol{\beta}}^m)$ give g_i and u_i values, respectively, for every $i \in \mathscr{D}_m$.

In order to construct restricted games of level 1, consider the ergodic structure under $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ for every $\mathscr{D}_m$ such that $\mathscr{D}_m \subseteq L_0$. Identify each recurrent class, say $\mathscr{R}_z$, in $\mathscr{D}_m$ and let $\mathscr{Z}_m$ be the set of recurrent classes in $\mathscr{D}_m$ under $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ for every $\mathscr{D}_m \subseteq L_0$. Since g_i is the same for every $i \in \mathscr{R}_z$, it will be denoted by $\bar{g}_{zm}$. Replace every recurrent class $\mathscr{R}_z$ with an absorbing aggregate state z. Define $\mathscr{T}_m$ as the set of transient states in $\mathscr{D}_m$ under $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ for $\mathscr{D}_m \subseteq L_0$, i.e., $\mathscr{T}_m = \mathscr{D}_m - \left(\bigcup_{z \in \mathscr{Z}_m} \mathscr{R}_z\right)$. The transient states in $\mathscr{T}_m$ are kept as they are. With the use of these aggregate and transient states, the restricted games to be solved at level 1 are defined as follows: For each maximal communicating class $\mathscr{D} \subseteq L_1$, a restricted game is constructed. The state space of the corresponding restricted game, to be denoted by $\bar{\mathscr{S}}$, is the union of $\mathscr{D}$, and the states accessible from $\mathscr{D}$. The latter would be some aggregate states in $\mathscr{Z}_m$ and/or some transient states in $\mathscr{T}_m$ such that $\mathscr{D}_m \subseteq L_0$. For each aggregate state z, abstract actions θ_{z1} and θ_{z2} are defined for the first and the second players, respectively. Then, action spaces of aggregate state z are $\bar{\mathscr{A}}_z = \{\theta_{z1}\}$ and $\bar{\mathscr{B}}_z = \{\theta_{z2}\}$. The corresponding payoff $\bar{r}_{z\theta_{z1}\theta_{z2}}$ is equal to $\bar{g}_{zm}$ for $z \in \mathscr{Z}_m$. For every state h in $\mathscr{D}$, the action spaces and the payoff amounts are kept the same as in the original stochastic game, i.e., $\bar{\mathscr{A}}_h = \mathscr{A}_h$, $\bar{\mathscr{B}}_h = \mathscr{B}_h$ and $\bar{r}_{hab} = r_{hab}$. For each transient state x in $\mathscr{T}_m$, which is included in $\bar{\mathscr{S}}$, $\hat{\boldsymbol{\alpha}}_x^m$ and $\boldsymbol{\beta}_x^m$ are fixed. The law of motion is given by transition matrix $\bar{P}$. For $z \in \mathscr{Z}_m$ such that $\mathscr{D}_m \subseteq L_0$, since action pair $(\theta_{z1}, \theta_{z2})$ is absorbing, $\bar{P}_{z\theta_{z1}\theta_{z2}z} = 1$. For a state x in $\mathscr{T}_m$ such that $\mathscr{D}_m \subseteq L_0$, the value of $\bar{r}_x$ is equal to $r_x(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ and

$$\bar{P}_{xl} = \begin{cases} \sum_{j \in \mathscr{R}_l} P_{xj}(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m) & \text{if } l \in \mathscr{Z}_m, \\ P_{xl}(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m) & \text{if } l \in \mathscr{T}_m. \end{cases}$$

For every $h \in \mathscr{D}$,

$$\bar{P}_{habl} = \begin{cases} \sum\limits_{j \in \mathscr{R}_l} P_{habj} & \text{if } l \in \mathscr{Z}_m \text{ such that } \mathscr{D}_m \subseteq L_0, \\ P_{habl} & \text{if } l \in \mathscr{T}_m \text{ such that } \mathscr{D}_m \subseteq L_0, \\ P_{habl} & \text{if } l \in \mathscr{D}. \end{cases}$$

Solutions of the restricted games constructed at level 1 give the best stationary strategies for the states in L_1. Note that best stationary strategies of the states in L_0 are obtained by the restricted games of level 0 and these strategies are kept fixed in the restricted games of level 1. In order to construct restricted games of level 2, every recurrent class $\mathscr{R}_z$ and every transient state x in L_1 are identified under the best stationary strategy pair of the restricted games of level 1. Each recurrent class is replaced with an aggregate state z and transient states are kept as they are. For each aggregate state z, abstract absorbing action θ_{z1}, θ_{z2} are defined and for each transient state the best stationary strategies found at level 1 are fixed. The procedure proceeds this way with the construction and solution of a restricted game for each maximal communicating class at every level.

Based on the best stationary strategies found for the restricted games at levels $n, n-1, \ldots, 0$, i.e., $(\hat{\boldsymbol{\alpha}}^y, \boldsymbol{\beta}^y)$ for each maximal communicating class $\mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n} L_d\right)$, a procedure to construct the restricted games of level $(n+1)$ is given below.

- *Identify the set of recurrent classes, $\mathscr{Z}_m$, and the set of transient states, $\mathscr{T}_m$, under $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ for every maximal communicating class $\mathscr{D}_m \subseteq L_n$.*
- *Replace each recurrent class $\mathscr{R}_z, z \in \mathscr{Z}_m$, such that $\mathscr{D}_m \subseteq L_n$, with an aggregate state z and define abstract absorbing actions θ_{z1}, θ_{z2}. Keep each transient state $x \in \mathscr{T}_m$, $\mathscr{D}_m \subseteq L_n$, as it is and fix its strategy pair as $(\hat{\boldsymbol{\alpha}}_x^m, \boldsymbol{\beta}_x^m)$.*
- *Define transition matrices and payoff values as follows:*
  - *For $z \in \mathscr{Z}_m$ such that $\mathscr{D}_m \subseteq L_n$, $\bar{r}_{z\theta_{z1}\theta_{z2}} = \bar{g}_{mz}$ and $\bar{P}_{z\theta_{z1}\theta_{z2}z} = 1$.*
  - *For $x \in \mathscr{T}_m$ such that $\mathscr{D}_m \subseteq L_n$, let $\bar{r}_x = r_x(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ and*

$$\bar{P}_{xl} = \begin{cases} \sum\limits_{j \in \mathscr{R}_l} P_{xj}(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m) & \text{if } l \in \mathscr{Z}_m \\ & \text{or } l \in \mathscr{Z}_y \text{ such that } \mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n-1} L_d\right), \\ P_{xl}(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m) & \text{if } l \in \mathscr{T}_m \\ & \text{or } l \in \mathscr{T}_y \text{ such that } \mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n-1} L_d\right). \end{cases}$$

  - *For every $z \in \mathscr{Z}_y$ and $x \in \mathscr{T}_y$ such that $\mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n-1} L_d\right)$, keep parameters the same as in the restricted games of level n.*
  - *In a restricted game defined for a maximal communicating class $\mathscr{D} \subseteq L_{n+1}$, for every $h \in \mathscr{D}$, let $\bar{\mathscr{A}}_h = \mathscr{A}_h$, $\bar{\mathscr{B}}_h = \mathscr{B}_h$ and $\bar{r}_{hab} = r_{hab}$ and*

$$\bar{P}_{habl} = \begin{cases} \sum_{j \in \mathscr{R}_l} P_{habj} & \text{if } l \in \mathscr{Z}_y \text{ such that } \mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n} L_d\right), \\ P_{habl} & \text{if } l \in \mathscr{T}_y \text{ such that } \mathscr{D}_y \subseteq \left(\bigcup_{d=0}^{n} L_d\right), \\ P_{habl} & \text{if } l \in \mathscr{D}. \end{cases}$$

Note that this procedure is employed for each restricted game at every level to obtain g_i and $(\hat{\boldsymbol{\alpha}}_i^*, \boldsymbol{\beta}_i^*)$ for all $i \in \mathscr{S}$. A similar one has to be employed to obtain u_i values and $(\boldsymbol{\alpha}_i^*, \hat{\boldsymbol{\beta}}_i^*)$. An explanation is not given for the latter problem, because the idea is the same as in the former one.

4.2 The decomposition algorithm

Based on the development in the previous section, the proposed decomposition algorithm is presented below.

Decomposition Algorithm
Step 1) Identify maximal communicating classes $\mathscr{D}_1, \ldots, \mathscr{D}_\Omega$.
Step 2) Identify levels of the maximal communicating classes. Let $n = 0$.
Step 3) Construct restricted games of level n and solve them for $(\hat{\boldsymbol{\alpha}}^m, \boldsymbol{\beta}^m)$ *and* $(\boldsymbol{\alpha}^m, \hat{\boldsymbol{\beta}}^m)$ *for each maximal communicating class* $\mathscr{D}_m \subseteq L_n$. *Let* $(\boldsymbol{\alpha}_i^*, \boldsymbol{\beta}_i^*) = (\boldsymbol{\alpha}_i^m, \boldsymbol{\beta}_i^m)$ *for every* $i \in \mathscr{D}_m$, $\mathscr{D}_m \subseteq L_n$.
Step 4) If $\left(\bigcup_{d=0}^{n} L_d\right) = \mathscr{S}$, *stop. Otherwise, increment n by 1 and go to step 3.*

A formal proof is given below to show that the decomposition algorithm works although it is immediately observed that this result is the consequence of proposition 1 and the independence of the restricted games from each other at every level.

Proposition 2. *Proposed decomposition algorithm gives the best stationary strategies for the undiscounted two-person zero-sum stochastic games.*

Proof: The proof is by induction for subproblem $\min_\beta \max_\alpha \phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $i \in \mathscr{S}$. The same arguments can also be used to give a proof for subproblem $\max_\alpha \min_\beta \phi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}), i \in \mathscr{S}$.

If the initial states are restricted to level 0, then *Problem 1* becomes equivalent to *Problem* 2_0. This reduction in *Problem 1* results from the definition of L_0, i.e., none of the states in $(\mathscr{S} - L_0)$ is accessible from the states in L_0. By definition of closed maximal communicating classes, the collection of the (independent) restricted games constructed at level 0 is equivalent to *Problem* 2_0. These restricted games are solved in the third step of the algorithm. From proposition 1, minimax part of *Problem* 2_0 gives stationary strategies $(\hat{\boldsymbol{\alpha}}_i^*, \boldsymbol{\beta}_i^*)$ for every $i \in L_0$, i.e., the algorithm works for $n = 0$.

By induction assumption, the stationary strategies $(\hat{\boldsymbol{\alpha}}_j^*, \boldsymbol{\beta}_j^*)$ for $j = \left(\bigcup_{d=0}^{n} L_d\right)$, and the corresponding g_j^* values are obtained by solving the restricted games constructed at levels $d = 1, \ldots, n$. Then, it has to be shown that the collection of the formulations for the restricted games constructed at level $(n+1)$ is equivalent to the minimax part of *Problem* 2_{n+1} where $\boldsymbol{\beta}_j$ and the corresponding maximizing $\boldsymbol{\alpha}$ strategy are fixed as $\boldsymbol{\beta}_j^*$ and $\hat{\boldsymbol{\alpha}}_j^*$, respectively, for every $j \in \left(\bigcup_{d=0}^{n} L_d\right)$.

Consider the long-run average expected payoff for an initial state, say i, in L_{n+1} under stationary strategies $(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ for every $j \in L_{n+1}$ and $(\hat{\boldsymbol{\alpha}}_j^*, \boldsymbol{\beta}_j^*)$ for every $j \in \left(\bigcup_{d=0}^{n} L_d\right)$:

$$\phi_i = \begin{cases} \sum_{d=0}^{n+1} \sum_{j \in L_d} \zeta_j^i r_j & \text{if } i \in \mathscr{T}_m \text{ such that } \mathscr{D}_m \subseteq L_{n+1}, \\ \sum_{j \in \mathscr{R}_z} \pi_j^z r_j & \text{if } i \in \mathscr{R}_z, z \in \mathscr{Z}_m \text{ such that } \mathscr{D}_m \subseteq L_{n+1}, \end{cases}$$

where π^z is the stationary probability vector given that the process is initially in recurrent class z, and $\mathscr{Z}_m$ ($\mathscr{T}_m$) is defined as before to denote the set of recurrent classes (transient states) in $\mathscr{D}_m$ under the considered strategies, ζ_j^i is the stationary probability of being in state j given that the initial state is i and r_j is the instantaneous payoff that depends on the strategies taken.

Let $\mathscr{Y}$ be the set of transient states in $\left(\bigcup_{d=0}^{n+1} L_d\right)$ under the specified stationary strategies. Note that $\mathscr{Y} = \bigcup_{d=0}^{n+1} \bigcup_{y \ni \mathscr{D}_y \subseteq L_d} \mathscr{T}_y$. Denote the transition probability matrix from states in $\mathscr{Y}$ to states in $\mathscr{Y}$ by $P^{\mathscr{Y}\mathscr{Y}}$. Let $P^{\mathscr{Y}z}$ be the transition probability matrix from states in $\mathscr{Y}$ to states in $\mathscr{R}_z$. If the process is initially in $\mathscr{Y}$, the first passage probabilities to $\mathscr{R}_z$ are given by $(I - P^{\mathscr{Y}\mathscr{Y}})^{-1} P^{\mathscr{Y}z}$. Let this matrix be called as F^z. Also, let $(I - P^{\mathscr{Y}\mathscr{Y}})^{-1}$ be denoted by Q. Then, $\boldsymbol{\zeta}^i$ is expressed in terms of $\boldsymbol{\pi}^z$ as follows: When $i \in \mathscr{R}_z$, $z \in \mathscr{Z}_m$ such that $\mathscr{D}_m \subseteq L_{n+1}$, the stationary probability ζ_j^i is equal to π_j^z for $j \in \mathscr{R}_z$, and zero otherwise. When $i \in \mathscr{T}_m$ such that $\mathscr{D}_m \subseteq L_{n+1}$, the stationary probability ζ_j^i is equal to $\pi_j^z \sum_{h \in \mathscr{R}_z} F_{ih}^z$ for $j \in \mathscr{R}_z$, $\mathscr{R}_z \subseteq \left(\bigcup_{d=0}^{n+1} L_d\right)$, and zero otherwise. Note that $\sum_{h \in \mathscr{R}_z} F_{ih}^z$ is the first passage probability to recurrent class z from initial state $i \in \mathscr{Y}$, and

$$\begin{aligned} \sum_{h \in \mathscr{R}_z} F_{ih}^z &= \sum_{h \in \mathscr{R}_z} \left(\sum_{j \in \mathscr{Y}} Q_{ij} P_{jh}^{\mathscr{Y}z} \right) \\ &= \sum_{j \in \mathscr{Y}} Q_{ij} \left(\sum_{h \in \mathscr{R}_z} P_{jh}^{\mathscr{Y}z} \right) \\ &= \sum_{j \in \mathscr{Y}} Q_{ij} \bar{P}_{jz}^{\mathscr{Y}z}, \quad \text{for } i \in \mathscr{Y}, \end{aligned}$$

where $\bar{P}_{jz}^{\mathscr{Y}}$ is the transition probability from transient state j to the aggregate state z. The relation above shows that $\sum_{h \in \mathscr{R}_z} F_{ih}^z$ is equal to the first passage probability from $i \in \mathscr{Y}$ to aggregate state z, say $\bar{F}_i^z$, in the corresponding restricted game. By making use of this observation, ϕ_i can be rewritten as follows:

If $i \in \mathscr{T}_m$ such that $\mathscr{D}_m \subseteq L_{n+1}$, then

$$\begin{aligned} \phi_i &= \sum_{d=0}^{n} \sum_{y \ni \mathscr{D}_y \subseteq L_d} \sum_{z \in \mathscr{Z}_y} \sum_{j \in \mathscr{R}_z} (\pi_j^z \bar{F}_i^z) r_j + \sum_{j \in L_{n+1}} \zeta_j^i r_j \\ &= \sum_{d=0}^{n} \sum_{y \ni \mathscr{D}_y \subseteq L_d} \sum_{z \in \mathscr{Z}_y} \bar{\zeta}_z^i \bar{g}_{zy} + \sum_{j \in \mathscr{D}_m} \zeta_j^i r_j, \end{aligned}$$

where $\bar{\zeta}_z^i$ is the stationary probability of being in aggregate state z given that the initial state is i in the restricted game of class $\mathscr{D}_m \subseteq L_{n+1}$. The second equality follows by $r_j = r_j(\hat{\boldsymbol{\alpha}}^*, \boldsymbol{\beta}^*)$ and $\bar{g}_{zm} = \sum_{j \in \mathscr{R}_z} \pi_j^z r_j$ and $\bar{\zeta}_z^i = \bar{F}_i^z$.

If $i \in \mathscr{R}_z$, $z \in \mathscr{Z}_m$ such that $\mathscr{D}_m \subseteq L_{n+1}$, then $\phi_i = \sum_{j \in \mathscr{R}_z} \pi_j^z r_j$.

Thus, ϕ_i's for $i \in L_{n+1}$, are also equal to the long-run average expected payoff amounts obtained from the collection of the restricted games where the recurrent classes under $(\hat{\boldsymbol{\alpha}}_j^*, \boldsymbol{\beta}_j^*)$, $j \in \left(\bigcup_{d=0}^n L_d\right)$, are replaced with aggregate states and $(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ is kept as it is for every $j \in L_{n+1}$. This proves that the collection of restricted games constructed at level $(n+1)$ gives $(\hat{\boldsymbol{\alpha}}_i^*, \boldsymbol{\beta}_i^*)$ for every $i \in L_{n+1}$. □

5 Conclusion

A decomposition procedure is proposed for undiscounted two-person zero-sum stochastic games based on the consideration of each maximal communicating class at only one of the disjoint levels of the state space. At the initial level, games restricted to absorbing maximal communicating classes are solved independently. Best stationary strategies of the states at each level $n \geq 1$ are determined by the best stationary strategies of the states at previous levels. Depending on the ergodic and/or data structure of the restricted games constructed at each level, one of the available algorithms may be used. In general, the use of NLP formulation due to Filar et al. [9] is suggested.

An extension of this decomposition approach can be used to solve undiscounted two-person nonzero-sum stochastic games. If the NLP formulation given in [9] is considered for the solution of restricted two-person nonzero-sum games, it is observed that the decomposition procedure should be used only for one pass to find stationary equilibrium strategies since this NLP formulation is not separable unlike *Problem 1*.

The motivation to devise a decomposition algorithm is to solve a stochastic game by dividing it into a number of smaller stochastic games. Especially, for the games with large state and/or action spaces the decomposition algorithm would make the solution procedure easier and faster as long as decomposition of the state space is not cumbersome. Also, when decomposition is used it is expected that the chance of finding better local optimal solutions is higher. In this study, problem in example 1(a) was solved using nonlinear programming solver MINOS. Considering the importance of initial points for nonlinear programming algorithms, both the NLP formulation for the original game and the decomposition procedure were employed with various initial solutions. When the initial point is not specified, MINOS assigns zero initially to each decision variable. Although this point was feasible for only one of the subproblems, the use of NLP for the whole problem gave the best solution. It also worked when the initial point was feasible. However, for specified infeasible initial points the value of the distance function obtained from the solution of NLP was too far from the best distance value. On the other hand, the decomposition procedure gave the best stationary strategies for each of those feasible and infeasible initial points. At this point, it should be noted that as the problem size gets larger, finding a feasible initial solution needs more effort. One other point to be noted is that there are iterative algorithms and LP formulations in the literature to solve games with certain properties.

Hence, instead of using NLP formulation for the original problem, these algorithms may be employed for the restricted games that have special structure, thus, further increasing the efficiency of the decomposition algorithm.

References

[1] Aumann RJ (1964) Mixed and behaviour strategies in infinite extensive games. Annals of Math. Studies 52
[2] Avşar ZM, Baykal-Gursoy M (1997) Two-person zero-sum communicating stochastic games. Technical Report, Industrial Engineering Department, Rutgers University
[3] Bather J (1973) Optimal decision procedures for finite Markov chains. Part III: General convex systems. Advanced Applied Probability 5:541–553
[4] Baykal-Gursoy M (1991) Two-person zero-sum stochastic games. Annals of Operations Research 28:135–152
[5] Federgruen A (1980) Successive approximation methods in undiscounted stochastic games. Operations Research 28:794–809
[6] Filar JA (1980) Algorithms for solving some undiscounted stochastic games. PhD thesis, University of Illinois at Chicago, Chicago, Illinois
[7] Filar JA, Raghavan TES (1980) Two remarks concerning two undiscounted stochastic games. Technical Report 392, John Hopkins University, Department of Mathematical Sciences
[8] Filar JA, Schultz TA (1986) Nonlinear programming and stationary strategies in stochastic games. Mathematical Programming 35:243–247
[9] Filar JA, Schultz TA, Thuijsman F, Vrieze DJ (1991) Nonlinear programming and stationary equilibria in stochastic games. Mathematical Programming 50:227–238
[10] Hoffman AJ, Karp RM (1966) On nonterminating stochastic games. Management Science 12:359–370
[11] Hordijk A, Kallenberg LCM (1981) Linear programming and Markov games I, II. In: Moeschlin O, Pallaschke D (eds) North Holland
[12] Parthasarathy T, Tijs SH, Vrieze OJ (1984) Stochastic games with state independent transitions and separable rewards. In: Hammer G, Pallaschke D (eds) Selected topics in OR and mathematical economics, Lecture Notes Series 226, Springer
[13] Raghavan TES, Filar JA (1991) Algorithms for stochastic games-A survey. ZOR-Methods and Models of Operations Research 35:437–472
[14] Raghavan TES, Tijs SH, Vrieze OJ (1985) On stochastic games with additive reward and transition structure. Journal of Optimization Theory and Applications 47:451–464
[15] Ross KW, Varadarajan R (1989) Markov decision processes with a sample path constraints: The communicating case. Oper. Res. 37:380–790
[16] Ross KW, Varadarajan R (1991) Multichain Markov decision processes with a sample path constraint: A decomposition approach. Mathematics of Operations Research: 195–207
[17] Van der Wal J (1980) Successive approximations for average reward markov games. International Journal of Game Theory 9:13–24
[18] Vrieze OJ (1981) Linear programming and undiscounted stochastic games. OR Spektrum 3:29–35
[19] Vrieze OJ, Tijs SH, Raghavan TES, Filar JA (1983) A finite algorithm for the switching controller stochastic game. OR Spektrum 5:15–24