STOCHASTIC GAMES, PRACTICAL MOTIVATION AND THE ORDERFIELD PROPERTY FOR SPECIAL CLASSES

O.J. VRIEZE
Maastricht University
Maastricht, The Netherlands

1. Introduction

Stochastic games concentrate on decision situations where at different time moments the players have to make a choice. The joint choices of all the players together have two implications. First, each player receives some reward, or loses some amount when this reward is negative. Second, the underlying system moves on along its trajectory. However, it is assumed that nature here plays a role in the sense that the transition might be the outcome of a random experiment, which might be dependent on the choices the players made. We only consider games where the decision moments are discrete points on a time axis and just for convenience we shall let these decision moments coincide with the set of positive natural numbers $\{1, 2, ...\}$. In stochastic games, perfect recall is assumed as well as complete information. That means that all the players know all the data of the game and at any future time moment all the players perfectly remember what has happened in the past.

The underlying system of a stochastic game is defined in terms of a state space S and the transitions in the course of the game are defined as moves from one state to another. Then, in any of the states players have so-called action sets, which might be state-dependent, and when the system arrives in a state each of the players has to choose, probably in a mixed way, an action out of his available set, etc.

Thus, when a stochastic game is played, each of the players is "rewarded" by a stream of immediate payoffs at the different decision moments.

In stochastic games the infinite horizon case is mostly studied. That is, the game never ends and there are a countable infinite number of decision moments. Though at first glance it looks as if studying a game of infinite

length is a huge task, there are nevertheless a few good reasons for it. From the practical viewpoint it is not always clear along how many steps a game will proceed. However, it is clear that the number of steps is immense. In such a situation a long-lasting game can very well be approximated by a game of infinite length. From a theoretical viewpoint there are several reasons for studying games of infinite length. One reason is that a finite game of finite length can be reformulated as a one-step game. Since we have a finite game tree there are just a finite number of strategies in this extended form game and by enumerating them, we can define a one-step finite game in normal form. Other reasons stem from interesting properties that stochastic games of infinite length exhibit, like stationarity in the discounted case and robustness of the solution in the undiscounted uniform approach. These interesting properties can be found throughout this book.

Stochastic games are motivated by many practical situations. We would now like to describe a few of them and we will shortly discuss how the model of a stochastic game suitably fits into the practical situation. But first of all we would like to emphasize the main tactical feature of a stochastic game, namely finding a balance between short-run "good" rewards and long-run "good" states. Being greedy during the beginning stages might seem advantageous. However, if the prize is that the system moves to states where the payoffs are relatively small and from which there is no escape, this starting profit will completely vanish in the long run. It is this tension that is characteristic of practical examples of stochastic games.

Pollution Game. Many industrial companies contribute to the pollution of the environment. Governmental bodies try to measure the damage caused by this pollution and in case of overpollution a tax will be raised. The companies have to decide every year whether to spend money for new technologies in order to reduce the pollution. Obviously, their market position is essential to their profit and spending much money on technologies reduces the advertisement and marketing budget with probably negative influence on the market position. This situation can be modeled as a stochastic game. The states of the system are a combination of the present market position and the pollution tax level. The actions of the companies are budget allocations to new technologies, advertisements, logistics, etc. The project is determined by the state (market share and tax costs) and the transitions are generally uncertain because they depend on consumer behavior and political tax rules.

Fishery Game. Fishing companies can try to catch as many fish as they can or they can catch moderately. If no fish are left in the ocean there is no next generation. Hence, it makes sense not to be too greedy. So every year the fishing companies have to decide about their quota and obviously the state space in this example is represented by the amount of fish in

the ocean at the start of the season. The uncertainty derives from weather conditions, which influence the reproduction rate, as well as from the fact that it is very hard to estimate the total number of fish.

Inspection Game. Big Brother is watching you. However, quite happily, it is not yet possible to watch all locations at once. For instance, when we think of an inspector who has to control incoming roads in a country in order to prevent drug smuggling, it might be clear that the inspector has to make choices of when to inspect and where to inspect. On the other hand, the smugglers face a similar problem in the sense that they have to guess when their smuggling route will be free.

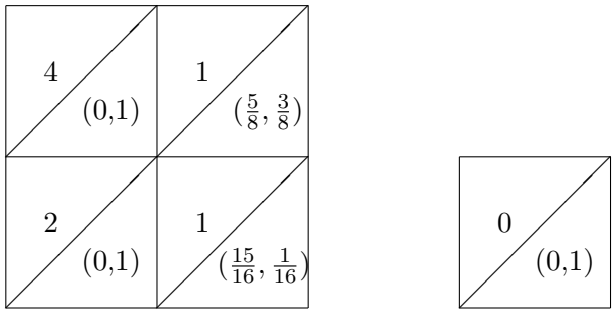
Salary Negotiations. Labor unions have to bargain with industrial companies about salaries and other working conditions. Typically, these processes go step by step, where at each step one or both of the parties will make a new offer. For the labor unions one of the available actions is a strike, obviously with the temporary drawback of a salary reduction. For the representatives of the industrial company there is always the threat of a strike, which evidently affects profits in the short run and perhaps market position in the long run as well. So again we see that both players have short-run incentives as well as long-run preferences. This negotiation situation can be perfectly modeled as a stochastic game. The uncertainty in this problem stems from the incomplete knowledge of the industrial representatives as to the union's willingness to strike. A second type of uncertainty comes from unpredictable market reaction to a strike.

2. The Orderfield Property

Whenever we face a problem that is described by finitely many parameters in a given domain, an interesting question concerns the search for a solution of the problem that lies in the given domain. For instance, a finite set of linear equations that has a solution can be solved by finitely many algebraic operations (addition, subtraction, multiplication and division), and thus it also has a solution in any field that contains all parameters of the system. Another example is the solution of a linear programming problem. If all parameters are from a fixed ordered field, the problem has a solution if and only if it has a solution in this fixed ordered field. A class of problems that are parameterized by finitely many elements from an arbitrary ordered field has the *orderfield property* if it has a solution in the same ordered field.

In particular, a class of game-theoretic problems that are parameterized by finitely many elements from an arbitrary ordered field has the orderfield property if it has a solution (e.g., minmax values, optimal strategies, or equilibrium strategies) in the same ordered field.

Stochastic games generally do not satisfy the orderfield property for any of the evaluation criteria. Consider the following zero-sum game, for which the value of the $\frac{1}{5}$-discounted game equals $\frac{1}{5}\sqrt{8}$.



State 1 State 2

The absence of the orderfield property is due to the nonlinearity of the Shapley equations:

$$\begin{cases} \gamma_\lambda(z) = \mathrm{Val}_{A(z)\times B(z)}[\lambda r(z,a,b) + (1-\lambda)\sum_{z'} p(z' \mid z,a,b)\gamma_\lambda(z')] \\ \text{for all } z \in S \end{cases}$$

which are equivalent to

$$\begin{cases} \max_{\alpha,\gamma} \sum_{z\in S} \gamma(z) \\ \text{subject to:} \\ \quad \gamma(z) \le \sum_a [\lambda r(z,a,b) + (1-\lambda)\sum_{z'} p(z' \mid z,a,b)\gamma(z')]\alpha(z,a), \\ \quad \text{for all } b \in B(z) \\ \quad \alpha(z,a) \ge 0, \ \text{ all } a \in A(z) \text{ and all } z \in S \\ \quad \sum_a \alpha(z,a) = 1, \ \text{ all } z \in S \end{cases}$$

The nonlinearity in the constraints is clear. When one wants to find classes of stochastic games for which the orderfield property holds, then mostly the defining conditions of these classes take care of a removal of this nonlinearity aspect.

In the rest of this chapter we will analyze several subclasses of stochastic games that give rise to the orderfield property. We will motivate these classes from applications.

3. Single-Controller Games

This class of games is motivated by the inspection model as mentioned above. Consider an inspector who has to control one out of finitely many sites every day. The violator tries to hide his illegal activity from the inspector and hopes to succeed at a site other than where the inspector is controlling. For this model the state space consists of the present site of the inspector. For both players the action sets consist of a collection of sites, one of which has to be chosen for the next day. One could easily build waiting days into this model. The payoff (i.e., cost) of the inspector is: "cost of travel + cost of inspection + cost of undetected violation - gain of an arrested criminal." For the violator different payoff functions could be relevant, for instance trying to minimize the probability of arrest or trying to maximize his gain in one way or another. Observe that in this model only the inspector determines the transitions and this observation has led to the study of the classes of games called *single-controller games*. Without loss of generality we may assume that player 2 is the controlling player. Then this class is defined as a standard stochastic game with the additional condition that $p(z' \mid z, a, b) = p(z' \mid z, \tilde{a}, b)$ for all $z', z, a, \tilde{a}, b$. So we can abbreviate the transitions to $p(z' \mid z, b)$, since they do not depend on the a-variable. Now one can easily check that this condition causes the nonlinear constraint in the above nonlinear program to become linear, namely

$$\gamma(z) \leq \sum_a \lambda r(z, a, b)\alpha(z, a) + (1 - \lambda) \sum_{z'} p(z' \mid z, b)\gamma(z').$$

So, for single-controller games, the Shapley equations yield a linear program and therefore the orderfield property holds for the discounted criterion in the zero-sum case.

Also for the limiting average criterion the orderfield property holds. This can be shown in two different ways. The first concerns a careful study of the limit process of the λ-discounted games when $\lambda \to 0$. It turns out that player 1 possesses a uniform discount optimal strategy (i.e., optimal for all λ in a neighborhood of 0) that is average optimal as well. Since the optimal strategy is of the data type it can be proved that the value is of that type as well. Further, along this approach, it follows that the solution of the limit discount equation is now a simple power series without fractional terms. The main statements in this spirit can be found in Parthasarathy and Raghavan [4] and in Filar and Raghavan [1]. The second approach for

the limiting average criterion of single-controller games is a straightforward formulation of a linear program that solves the game, namely

$$\begin{cases} \max\limits_{\gamma,\nu,\alpha} \sum\limits_{z} \gamma(z) \\ \text{subject to:} \\ \quad \gamma(z) \le \sum\limits_{z'} p(z' \mid z,b)\gamma(z'), \quad \text{all } z,b \\ \quad \gamma(z) + \nu(z) \le \sum\limits_{a} r(z,a,b)\alpha(z,a) + \sum\limits_{z'} p(z' \mid z,b)\nu(z'), \quad \text{all } z,b \\ \quad \alpha(z,a) \ge 0, \ \text{ all } a,z \\ \quad \sum\limits_{a} \alpha(z,a) = 1, \ \text{ all } z \end{cases}$$

The details of this linear program can be found in Vrieze [7].

For non-zero-sum single-controller games we can derive similar results. Again, different approaches can be found in the literature. We mention the approach of Nowak and Raghavan [3], who base their analysis on the bimatrix game constituted by the pure stationary strategies of the players. A second approach follows straightforwardly from the sufficiency condition at the end of this chapter.

An extension of the single-controller game is the so-called switching control game. In a switching control game in every state only one of the players controls the transitions. However, this is not necessarily the same player. So $S = S_1 \cup S_2$, with transitions $p(z' \mid z,a)$ for $z \in S_1$ and $p(z' \mid z,b)$ for $z \in S_2$. Examples of switching control games can be found in political situations where two parties dominate the scene, as in the U.S. Each party's chances of delivering the next president can be assumed to depend merely on the behavior and capability of the current president. So if the state space reflects the president's political party we get a switching control game. For switching control games it can be shown that the orderfield property holds for the zero-sum version both for the discounted criterion and for the limiting average criterion. In both cases the proof can be given with the aid of an iterative procedure. Each iteration solves an auxiliary one-player (either player 1 or player 2) control game. The auxiliary single-controller game is derived from the solution of the previous iteration by fixing the mixed actions of one of the controlling players in all states that he controls. The outcome of this single-controller auxiliary game (that obeys the orderfield property) serves as input for the next step of the procedure. It can be shown that this procedure reaches the solution of the game after finitely many iterations, thus demonstrating the orderfield property. The relevant facts can be found in Vrieze [8] and Vrieze et al. [9].

4. SER-SIT Games

It is conceivable that we have a decision situation where the transitions depend only on the present actions and not on the present state. For instance, in the pollution example mentioned in the introduction, the capability of the government to measure an abundant pollution obviously depends on the emission of pollutants only in the current year and not on past emissions. So, if in that example the state space is the tax level, then we arrive at a State Independent Transition (SIT) game. Notationally, the transition can be given as $p(z' \mid a, b)$.

Further, it is conceivable that the rewards of action combinations can be given as the sum of a term depending on the actions and of a term depending on the state. So, $r^i(z, a, b) = r^i(z) + r^i(a, b)$ for $i = 1, 2$. This feature is called the Separable Reward (SER) property. Again, referring to the pollution game, $r^i(z)$ denotes the state-dependent tax level and $r^i(a, b)$ denotes the profit of the companies, besides the tax obligations. We tacitly assume that the tax level does not influence market behavior, which only reacts to the marketing and advertising of the companies.

For SER-SIT games the action sets for both players are state-independent, so we can speak of action sets A and B. It is straightforward to show that the solution of SER-SIT games for the zero-sum version is given by

$$\gamma_\lambda(z) = \lambda r(z) + \nu_\lambda \quad (\lambda \geq 0),$$

where

$$\nu_\lambda = \mathrm{Val}_{A \times B}[r(a, b) + (1 - \lambda) \sum_{z'} p(z' \mid a, b) r(z')].$$

In this characterization $\lambda = 0$ yields the limiting average solution. Further, optimal strategies can be found by implementing a stationary strategy that subscribes an optimal action of the above matrix game in every state. So for SER-SIT games in every state the same action can be chosen which gives rise to a myopic strategy.

For the non-zero-sum version an analogous approach can be given, resulting in the same conclusion.

Obviously, SER-SIT games have the orderfield property, since matrix (and bimatrix) games have this property. As a last remark on SER-SIT games we mention that both the properties SER and SIT are independently needed for the orderfield property. If one of them does not hold, examples can be constructed that fail the orderfield property.

5. AR-AT Games

Additive Reward and Additive Transition (AR-AT) games concern the situations where the influence of the players can be added up. For instance, if

one recalls the fishery game discussed in the introduction, then the reproductive capability of the fish in the ocean is linear in the amount of fish. Hence, both fishing companies contribute negatively in an additive way to the supply of fish for the following year. So, if the state is represented by the amount of fish in the ocean we see that the transitions are additive with respect to the players, so

$$p(z' \mid z, a, b) = p(z' \mid z, a) + p(z' \mid z, b).$$

For this example the same additivity assumption holds for the rewards. If we assume the price of fish at the market to be independent of the actions of the players (i.e., the amount of fish they catch), then a player's payoff just depends on his own quota and his own fishing costs like equipment, salaries, etc. So we have

$$r^i(z, a, b) = r^i(z, a) + r^i(z, b),$$

for $i = 1, 2$, where for this example $r^1(z, b) = 0$ and $r^2(z, a) = 0$. For an AR-AT game the Shapley equations reduce to

$$\gamma_\lambda(z) = \mathrm{Val}_{A(z)\times B(z)}[\lambda r(z, a) + (1 - \lambda) \sum_{z\prime} p(z' \mid z, a)\gamma_\lambda(z') +$$

$$+\lambda r(z, b) + (1 - \lambda) \sum_{z\prime} p(z' \mid z, b)\gamma_\lambda(z')]$$

for all $z \in S$.

So we have to solve a matrix game for every state, where the payoff is the sum of a term dependent on action a and a term dependent on action b. But then it is easy to see that both players have pure optimal actions. Hence the orderfield property holds.

When λ tends to 0, obviously some pure optimal action for the λ-discounted game repeats itself infinitely often, since there are only finitely many candidates for it. Then it can be deduced that such an action is uniformly discount optimal and limiting average optimal as well, showing the orderfield property for the average criterion. For further reference to this class of games see Raghavan et al. [6].

Surprisingly, for SER-SIT games the orderfield property does not hold for the non-zero-sum version.

6. A Sufficiency Theorem

Until now there has been no known characterization for the class of games for which the orderfield property holds. In the eighties this topic got a lot of attention but a complete statement was never found.

The reason why this topic got a lot of attention derives from a computational insight. For a class of games without the orderfield property one cannot expect to compute an exact solution for a generic instance of this class. When nonlinear equations have to be solved, say by a suitable computer program, then generally the solution can only be approximated. For the discounted criterion this does no harm, since the value and (nearly) stationary strategies are continuous in the discount factor. However, especially for the limiting average case we might find problems, since the limiting average payoff is not a continuous function over the space of stationary strategies. So, slightly perturbed strategies might cause big changes in the payoffs. For games with the orderfield property we might expect to be able to find an exact solution, since we expect that there should be an algorithm with only finitely many multiplications or divisions in order to find a solution.

We now present a theorem that states sufficient conditions for a stochastic game to possess the orderfield property. It can be shown that all of the known results with respect to the orderfield property can be deduced either straightforwardly or indirectly from this theorem.

Take for any $z \in S$ a subset $\tilde{A}(z)$. Then the set $\underset{z\in S}{\times} \tilde{A}(z)$ can be interpreted as a set of pure stationary strategies for player 1. The same can be done for player 2 with subsets $\tilde{B}(z) \subseteq B(z)$.

Now consider the following maps F and G which are defined on these sets of pure stationary strategies or equivalently on collections $\left\{\tilde{A}(z) \mid z \in S\right\}$ respectively $\left\{\tilde{B}(z) \mid z \in S\right\}$:

$$F\left(\underset{z\in S}{\times} \tilde{A}(z)\right) := \left\{\beta \mid \text{all } \alpha \in \underset{z\in S}{\times} \tilde{A}(z) \text{ is a pure best answer against } \beta\right\}$$

and

$$G\left(\underset{z\in S}{\times} \tilde{B}(z)\right) := \left\{\alpha \mid \text{ all } \beta \in \underset{z\in S}{\times} \tilde{B}(z) \text{ is a pure best answer against } \alpha\right\}.$$

The carrier of a mixed action $\alpha(z)$ (denoted as $\mathrm{car}(\alpha(z))$) in a state $z \in S$ is defined as $\{a \mid \alpha(z,a) > 0\}$ and the carrier of a mixed action $\beta(z)$ is defined analogously.

The following theorem holds for any criterion and a proof can be found in Filar and Vrieze [2].

Theorem 1 *The pair of stationary strategies (α, β) is an equilibrium point if and only if*

$$\beta \in F\left(\underset{z\in S}{\times} car\left(\alpha(z)\right)\right) \text{ and } \alpha \in G\left(\underset{z\in S}{\times} car\left(\beta(z)\right)\right).$$

The proof is based on the observation that a player only makes use of a pure stationary strategy in his best response against a stationary strategy, when such a pure stationary strategy is a best response itself.

Now we can state our sufficiency theorem.

Theorem 2 *When for all* $\underset{z\in S}{\times}\tilde{A}(z)$ *and all* $\underset{z\in S}{\times}\tilde{B}(z)$ *with* $\tilde{A}(z)\subseteq A(z)$ *and* $\tilde{B}(z)\subseteq B(z),\forall z\in S$, *it holds that* $F\left(\underset{z\in S}{\times}\tilde{A}(z)\right)$ *as well as* $G\left(\underset{z\in S}{\times}\tilde{B}(z)\right)$ *can be written as a finite sum of polytopes with extreme points that satisfy the orderfield property, then the orderfield property holds for the stochastic game as well, provided that solutions do exist.*

This theorem has a general application range. It can be applied to zero-sum as well as to non-zero-sum. We will not give a rigorous proof but the following reasoning might provide the reader with an insight into the idea behind the proof.

Suppose we have a stochastic game for which the sufficiency theorem holds. Let (α,β) form an equilibrium point. By the above characterization of equilibrium points we see that all pure strategies belonging to $\underset{z\in S}{\times}\operatorname{car}(\beta(z))$ are best responses to α. Hence $\beta\in F\left(\underset{z\in S}{\times}\operatorname{car}(\alpha(z))\right)$ and $\alpha\in G\left(\underset{z\in S}{\times}\operatorname{car}(\beta(z))\right)$.

Suppose that the game has rational data. Then by the sufficiency theorem we get that α is an element of a polytope with rational extreme points and likewise β. Now we claim that there exists an element $\tilde{\alpha}$ in this polytope with rational components for which $\operatorname{car}(\alpha(z))=\operatorname{car}(\tilde{\alpha}(z))$ for all $z\in S$. Likewise we claim that there exists an element $\tilde{\beta}$ of the polytope containing β with rational components such that $\operatorname{car}(\beta(z))=\operatorname{car}\left(\tilde{\beta}(z)\right)$ for all $z\in S$. It then follows that $\left(\tilde{\alpha},\tilde{\beta}\right)$ forms an equilibrium point with rational components.

This theorem can be used in proving the orderfield property for all the known classes. However, it is not clear to us whether indeed this sufficiency condition is necessary for a game to possess the orderfield property.

References

1. Filar, J.A. and Raghavan, T.E.S. (1984) A matrix game solution of the single-controller stochastic game, *Mathematics of Operations Research* **9**, 356–362.
2. Filar, J.A. and Vrieze, O.J. (1997) *Competitive Markov Decision Processes*, Springer-Verlag, New York.
3. Nowak, A. and Raghavan, T.E.S. (1989) A finite-step algorithm via a bimatrix game to a single controller non-zero-sum stochastic game, *Mathematical Programming* **59**, 249–259.

4. Parthasarathy, T. and Raghavan, T.E.S. (1981) An orderfield property for stochastic games when one player controls transition probabilities, *Journal of Optimization Theory and Applications* **33**, 375–392.
5. Parthasarathy, T., Tijs, S.H. and Vrieze, O.J. (1984) Stochastic games with state independent transitions and separable rewards, in G. Hammer and D. Pallaschke (eds.), *Selected Topics in OR and Mathematical Economics*, Lecture Notes Series 226, Springer-Verlag, Berlin, pp. 262–271.
6. Raghavan, T.E.S., Tijs, S.H. and Vrieze, O.J. (1985) On stochastic games with additive reward and transition structure, *Journal of Optimization Theory and Applications* **47**, 451–464.
7. Vrieze, O.J. (1981) Linear programming and undiscounted stochastic games, *OR Spektrum* **3**, 29–35.
8. Vrieze, O.J. (1987) Stochastic games with finite state and actions spaces, CWI-Tract 33, CWI, Amsterdam.
9. Vrieze, O.J., Tijs, S.H., Raghavan, T.E.S. and Filar, J.A. (1983) A finite algorithm for the switching controller stochastic game, *OR Spektrum* **5**, 15–24.