The Annals of Mathematical Statistics
1971, Vol. 42, No. 6, 1930–1935

NONCOOPERATIVE STOCHASTIC GAMES[1]

BY MATTHEW J. SOBEL
Yale University

1. Summary. We introduce a sequential competitive decision process that is a generalization of noncooperative finite games and of two-person zero-sum stochastic games (hence, of Markovian decision processes). We prove the existence of equilibrium points under criteria of discounted gain and of average gain.

Two person zero-sum stochastic games and noncooperative finite games were introduced in elegant papers by Shapley [22] and Nash [16], [17]. Shapley's work prompted a series of papers [1], [4], [5], [10], [11], [12], [14], [18], [26] concerned with the existence of minimax solutions and algorithms for their computation. Even for the two-person zero-sum case, no finite algorithm yet exists. Nash's papers led to a sizeable literature in both mathematics and economics. Mills' [15] work, for example, is related to our characterization of equilibrium points in Section 4.

Noncooperative stochastic games may yield fruitful models for several phenomena in the social sciences. Theories of economic markets, for example, have increasingly sought to encompass sequential economic decision processes. Some recent research in social psychology has taken an analogous direction [19], [25].

I became aware of recent work by Rogers [20] shortly after completing this paper. His results and ours nearly coincide with our Theorem 2 being slightly stronger than the comparable results in his paper. The basic difference between the papers is that Rogers relies on the Kakutani fixed point theorem whereas we use Brouwer's theorem. Our arguments are somewhat simpler as a consequence.

2. Preliminaries. A noncooperative stochastic game Γ is a sequence $\gamma_1, \gamma_2, \cdots$, where, for each t, $\gamma_t \in \{\Gamma_1, \cdots, \Gamma_n\}$. We call $S = \{1, \cdots, n\}$ the *state space*. For each *state* $s \in S$, Γ_s is the following N-person non-zero-sum noncooperative game. The set of actions available to the ith player is $A_s^i = \{1, \cdots, K_s^i\}$ with $K_s^i \geqq 1$. A noncooperative stochastic game is *finite* if $\sum_{s \in S} \sum_{i \in \Omega} K_s^i < \infty$ where $\Omega = \{1, \cdots, N\}$. The reward to player i is $r_s^i(a)$ when the actions of the players is given by $a = (a_1, \cdots, a_N) \in B_s = \mathbf{X}_{k \in \Omega} A_s^k$. The assumption $K_s^i = K$ so $A_s^i = A$ and $B_s = B$ for all s and i entails no loss of generality and is henceforth made.

The conditional probability that γ_{t+1} is Γ_j given that γ_t is Γ_s, that actions $a \in B$ were taken in γ_t, and given the observed states and actions taken at $\gamma_1, \cdots, \gamma_{t-1}$, is assumed to be a function $q_{sj}(a)$ depending only on s, j, and a. Shapley [22] has distinguished the "terminating" case (i) $\sum_j q_{sj}(a) < 1$ for each s and a, from the "non-terminating" case (ii) $\sum_j q_{sj}(a) = 1$ for each s and a. Probabilistically, (i) is a special case of (ii) having an absorbing state which, with probability one, is

Received November 21, 1969; revised May 12, 1971.

[1] Partially supported by National Science Foundation Grant GK–13757 and by a Yale University Junior Faculty Fellowship in Social Science.

entered finitely whatever the initial state and actions. Thus, we assume (ii) and use the following means to separate the terminating and non-terminating cases.

Let $\beta_s^i(a)$ satisfy $0 < \beta_s^i(a) < 1$ for all $a \in B$, $s \in S$, $i \in \Omega$. Then $\beta_s^i(a)$ can be interpreted as the present value to player i of receiving a unit reward in the subsequent state occupied if the present state is s in which the players take actions a. Ordinarily, for each $i \in \Omega$, $\beta_s^i(a)$ will be constant for all s and a. The generality is useful, however, because it extends Theorems 1 and 3 to noncooperative stochastic games based on *Markov renewal programs* ([6] and its references) rather than discrete-time Markov decision processes. It can be shown that a discounted Markov renewal program is equivalent to a discounted discrete-time model whose discount factor is a function of the state occupied and of the action taken.

For an outcome of Γ, let X_t^i and α_t^i be player i's reward and discount factor in γ_t; define $\alpha_0^i = 1$, $i \in \Omega$. Then, using $V^i = \sum_{t=1}^{\infty} X_t^i \prod_{\tau=1}^{t} \alpha_\tau^i$, ($G^i = \liminf_{T\to\infty} T^{-1} \sum_{t=1}^{T} X_t^i$) is tantamount to the terminating case (non-terminating case).

A stationary policy δ_i for player i is a sequence of probability vectors $(D_s^i, s \in S)$ such that D_s^i is a randomized strategy for player i in γ_t if $\gamma_t = \Gamma_s$. Thus, $D_s^i = (D_{sk}^i)$ is a K-vector with $D_{sk}^i \geqq 0$ and $\sum_{k \in A} D_{sk}^i = 1$. The set F of feasible D_s^i is the $K-1$ dimensional unit simplex. Then $\pi = \mathbf{X}_{s \in S} F$ is the set of all δ_i and $\boldsymbol{\pi} = \mathbf{X}_{i \in \Omega} \pi$ is the set of N-tuples $\delta = (\delta_1, \cdots, \delta_N)$ of all the players' stationary policies. *No class of policies larger than $\boldsymbol{\pi}$ is considered in this paper.* The players' stationary policies excluding those of player i is $\pi^- = \boldsymbol{\pi} - \pi$ and for $\delta \in \boldsymbol{\pi}$ we use the (abused) notation $\delta = (\delta_i, \delta^{-i}) \in \pi \mathbf{X} \pi^-$. Finally, let δ_{sk}^i denote the modification of δ whereby $D_{sk}^i = 1$ and $D_{sj}^i = 0$ if $j \neq k$.

Suppose (only for this paragraph) that the players, except for the ith, adhere to $\delta^{-i} \in \pi^-$. It is essential to our results to observe that player i confronts a Markovian decision process. For an arbitrary (possibly non-stationary) measurable sequential policy let $V_s^i = E(V^i \mid \gamma_1 = \Gamma_s)$ and $G_s^i = E(G^i \mid \gamma_1 = \Gamma_s)$. With either criterion in a finite game it follows from Blackwell [2] or Derman [9] in conjunction with [8] that player i experiences no loss of optimality from restricting his policy to π.

For any $\delta \in \boldsymbol{\pi}$, let $v_s^i(\delta)$ and $g_s^i(\delta)$ make explicit the dependence of V_s^i and G_s^i on the policies. Let v_δ^i and g_δ^i be the associated n-vectors of $v_s^i(\delta)$ and $g_s^i(\delta)$, $s \in S$.

DEFINITION 1. A stationary policy $\delta \in \boldsymbol{\pi}$ is an undiscounted equilibrium point (UEP) iff

$$g_\delta^i = \max \{g_{(\rho,\delta^{-i})}^i \mid \rho \in \pi\}, \qquad i \in \Omega. \tag{2.1}$$

DEFINITION 2. A stationary policy $\delta \in \boldsymbol{\pi}$ is a discounted equilibrium point (DEP) iff

$$v_\delta^i = \max \{v_{(\rho,\delta^{-i})}^i \mid \rho \in \pi\}, \qquad i \in \Omega. \tag{2.2}$$

If $n = 1$ then (2.1) and (2.2) reduce to the definition [17] of a Nash equilibrium point. If $N = 2$ and the rewards are zero-sum (constant-sum) then the definitions characterize a minimax solution.

Recent workers in the field of dynamic programming have obtained various expansions for quantities such as G_s^i, an expected average gain per game played. Thus motivated, we shall define a subset of the UEPs. For $\delta \in \pi$ given by $\{D_{sk}^i\}$ let $\mathbf{a}_s^\delta$ be the actions taken by the players in state s;

$$P\{\mathbf{a}_s^\delta = (a_1, \cdots, a_N)\} = \prod_{j \in \Omega} D_{sa_j}^j. \tag{2.3}$$

Let P_δ be the stochastic matrix with elements $p_{su}(\delta) = Eq_{su}(\mathbf{a}_s^\delta)$ and for $i \in \Omega$ let r_δ^i be the n-vector with components $r_s^i[\delta] = E_\delta r_s^i(\mathbf{a}_s^\delta)$ $s \in S$; $r_s^i[\delta]$ is the expected gain under δ to player i each time that Γ_s is played. It is well known that for each P_δ there is a unique stochastic matrix P_δ^* such that $(1/T)\sum_{t=0}^{T-1} P_\delta^t \to P_\delta^*$. Then it follows from Blackwell [2] that for each $\delta \in \pi$ and $i \in \Omega$ there is a unique pair of n-vectors (g_δ^i, w_δ^i) such that

$$P_\delta g_\delta^i = g_\delta^i, \qquad P_\delta^* g_\delta^i = P_\delta^* r_\delta^i \tag{2.4a}$$

$$r_\delta^i + P_\delta w_\delta^i = w_\delta^i + g_\delta^i \tag{2.4b}$$

$$P_\delta^* w_\delta^i = 0. \tag{2.4c}$$

DEFINITION 3. A stationary policy $\delta \in \pi$ is a (g, w)-equilibrium point ((g, w)EP) iff (2.1) and

$$w_\delta^i = \max\{w_{(\rho, \delta^{-i})}^i \mid \rho \in \pi\}, \qquad i \in \Omega. \tag{2.5}$$

3. Existence.

THEOREM 1. *Every finite noncooperative stochastic game has a* DEP.

PROOF. Following Nash [17], we construct a continuous mapping $\tau: \pi \to \pi$ whose fixed points are DEPs and conversely. Let

$$p_{su}^i(\delta) = E\beta_s^i(\mathbf{a}_s) q_{su}(\mathbf{a}_s^\delta) \tag{3.1}$$

be the expected discounted transition probability, and let P_δ^i be the matrix of $p_{su}^i(\delta)$. Then

$$v_\delta^i = \sum_{t=0}^{\infty} (P_\delta^i)^t r_\delta^i = (I - P_\delta^i)^{-1} r_\delta^i \tag{3.2}$$

as geometric convergence follows from (3.1) and $0 < \beta_s^i(a) < 1$ for all s, a, and i.

Let $v_{sk}^i(\delta) = r_s^i[\delta_{sk}^i] + \sum_u p_{su}^i(\delta_{sk}^i) v_s^i(\delta)$ and $\phi_{sk}^i(\delta) = \max\{0, v_{sk}^i(\delta) - v_s^i(\delta)\}$. Then $\tau: \pi \to \pi$ with $\bar{\delta} = \tau(\delta)$ is defined by

$$\bar{D}_{sk}^i = (D_{sk}^i + \phi_{sk}^i(\delta))/(1 + \sum_{j \in A} \phi_{kj}^i(\delta)), \qquad k \in A, s \in S, i \in \Omega. \tag{3.3}$$

The continuity on π of $v_s^i(\cdot)$ and $v_{sk}^i(\cdot)$, hence of $\phi_{sk}^i(\cdot)$ and τ, follows from (3.2) and the convexity of π. The existence of a fixed point of τ is given by the Brouwer fixed point theorem.

It remains to show that the set of fixed points of τ coincides with the set of DEPs. If δ is a DEP then Howard's [13] policy improvement algorithm implies $\phi_{sk}^i(\delta) = 0$ for all s, k, i, so δ is a fixed point. If δ is a fixed point then (3.3) yields $\phi_{sk}^i = D_{sk}^i$

$\sum_j \phi^i_{sj}$ for all s, k, and i. Thus $\phi^i_{sk}(\delta) = 0$ if $D^i_{sk} = 0$. Suppose $\sum_u \sum_j \phi^i_{uj}(\delta) > 0$ for some i. Then $D^i_{sk} = \phi^i_{sk}/\sum_j \phi^i_{sj}$ for all s and k such that $D^i_{sk} > 0$. However, for some $s \in S$ and $j \in A$ there is $D^i_{sj} > 0$ and $\phi^i_{sj} = 0$:

$$v_s{}^i(\delta) = \sum_{k:D^i_{sk}>0} D^i_{sk} v^i_{sk}(\delta) \geqq \min\{v^i_{sk}(\delta) \mid k \in A \quad \text{and} \quad D^i_{sk} > 0\} = v^i_{sj}(\delta), \quad \text{say,}$$

so $\phi^i_{sj}(\delta) = 0 < D^i_{sj}$. Therefore, $\phi^i_{sk}(\delta) = 0$ for all s, k, and i which implies by [13] that δ is a DEP. □

Unlike the case $N = 1$, i.e. Markovian decision process, generally we cannot assert the existence of a DEP in pure strategies. This precludes exploiting Theorem 1 to prove the existence of UEPs or (g, w)EPs. Moreover, a UEP need not exist; Gillette [11] has a counterexample.[2] Let **B** denote the finite set of pure (unrandomized) policies in π.

THEOREM 2. *A finite noncooperative stochastic game has a* (g, w)EP *if for every* $\delta \in \mathbf{B}$, P_δ *has exactly one communicating class of states.*

Theorem 2 does not require that the communicating classes coincide. However, it does rule out games with several disjoint classes for the same $\delta \in \mathbf{B}$—a property of Gillette's counterexample. On the other hand, a simple modification of the reward structure in Gillette's example ($a^3_{k1} = 0$ on page 185 of [11]) yields a game with a UEP but not meeting the condition of Theorem 2.

PROOF. The method of proof is similar to that of Theorem 1 but depends on Veinott's [24] algorithm and a result by Denardo [6]. Again, following Nash [17], we construct a continuous mapping $\tau: \pi \to \pi$. Let $\phi^i_{sk}(\delta) = a^i_{sk} + b^i_{sk} + c^i_{sk}$ where $a^i_{sk} = \max\{0, \sum_u p_{su}(\delta^i_{sk}) g_u{}^i(\delta) - g_s{}^i(\delta)\}$; $b^i_{sk} = 0$ if $\sum_s \sum_k a^i_{sk} > 0$ and $b^i_{sk} = \max\{0, r_s{}^i[\delta^i_{sk}] + \sum_u p_{su}(\delta^i_{sk}) w_u{}^i(\delta) - g_s{}^i(\delta) - w_s{}^i(\delta)\}$ if $\sum_s \sum_k a^i_{sk} = 0$; $c^i_{sk} = 0$ if $\sum_s \sum_k b^i_{sk} > 0$ and $c^i_{sk} = \max\{0, \sum_u p_{us}(\delta^i_{sk}) z_u{}^i(\delta) - w_s{}^i(\delta) - z_s{}^i(\delta)\}$ if $\sum_s \sum_k b^i_{sk} = 0$ where by [2], [24] for each $\delta \in \pi$, $i \in \Omega$, z^i_δ is the unique solution to

$$z_\delta{}^i + w_\delta{}^i = P_\delta z_\delta{}^i, \qquad P_\delta{}^* z_\delta{}^i = 0; \tag{3.4}$$

$w_s{}^i(\delta)$ and $z_s{}^i(\delta)$, $s \in S$, are the components of $w_\delta{}^i$ and $z_\delta{}^i$.

Again, let (3.3) define τ. Continuity of τ on π will be shown below. Then Brouwer's theorem asserts the existence of a fixed point. To show that a fixed point δ is an equilibrium point, we have as in Theorem 1 that $\phi^i_{sk} = 0$ for all s, k, and i. Therefore, Veinott's algorithm [24] implies δ is a (g, w)-equilibrium point. Conversely, if δ is an equilibrium point then all $\phi^i_{sk} = 0$ from [6] so δ is a fixed point.

It remains to show continuity on π of $g_s{}^i(\cdot)$, $w_s{}^i(\cdot)$, and $z_s{}^i(\cdot)$. Suppose $g_s{}^i(\delta)$, P_δ, and $P_\delta{}^*$ are continuous on π. Then continuity of $w_s{}^i(\cdot)$ and $z_s{}^i(\cdot)$ follows from (2.4b, c), (3.4), and [2]. For continuity of $g_s{}^i(\cdot)$, it suffices to show continuity of P_δ and $P_\delta{}^*$ and to verify that P_δ has one communicating class of states.

---

[2] I am grateful to Eric V. Denardo for pointing out the example and for noticing an error in an earlier version of this paper.

π is a convex set with finite extreme point set B. Therefore, $\delta \in \pi$ implies $P_\delta = \sum_{u \in B} \alpha_u P_u$ with $\alpha_u \geqq 0$ and $\sum_u \alpha_u = 1$. Hence, the assumption of this theorem and Theorem 1 in Schweitzer [21] ensure the needed class structure. The continuity properties are an easy consequence of the class structure. □

4. Necessary and sufficient conditions. Characterizations of equilibrium points are useful to infer qualitative properties of solutions to particular games. Also, they may facilitate computation of an equilibrium point. For the special case of a (static) noncooperative game, i.e. $|S| = 1$, the following results were exploited in [23] to develop a finite algorithm.

A characterization of the DEPs is given by

THEOREM 3. *δ is a* DEP iff *$\{D^i_{sk}\}$ is part of a solution to*

$$\lambda^i_{sk} \geqq 0, \qquad D^i_{sk} \geqq 0, \qquad k \in A, s \in S, i \in \Omega; \tag{4.1a}$$

$$\sum_{k \in A} D^i_{sk} = 1, \qquad s \in S, i \in \Omega, \tag{4.1b}$$

$$v_s{}^i = \lambda^i_{sk} + r_s{}^i[\delta^i_{sk}] + \sum_{u \in S} p^i_{su}(\delta^i_{sk}) v_u{}^i, \qquad k \in A, s \in S, i \in \Omega, \tag{4.2}$$

$$\sum_{i \in \Omega} \sum_{s \in S} \sum_{k \in A} \lambda^i_{sk} D^i_{sk} = 0. \tag{4.3}$$

PROOF. (λ, D, v) satisfies (4.1)–(4.3) iff D generates $\delta \in \pi$ and δ attains

$$v_s{}^i = \max_{k \in A} \{r_s{}^i[\delta^i_{sk}] + \sum_u p^i_{su}(\delta^i_{sk}) v_u{}^i\}, \qquad s \in S, i \in \Omega, \tag{4.4}$$

which by [13] is necessary and sufficient for (2.2). □

A characterization of UEPs analogous to Theorem 3 is

THEOREM 4. *δ is a* UEP iff *$\{D^i_{sk}\}$ is part of a solution to*

$$\lambda^i_{sk} \geqq 0, \qquad \mu^i_{sk} \geqq 0, \qquad D^i_{sk} \geqq 0, \qquad k \in A, s \in S, i \in \Omega, \tag{4.5a}$$

$$\sum_{k \in A} D^i_{sk} = 1, \qquad s \in S, i \in \Omega, \tag{4.5b}$$

$$\xi_s{}^i + \omega_s{}^i = \lambda^i_{sk} + r_s{}^i[\delta^i_{sk}] + \sum_u p_{su}(\delta^i_{sk}) \omega_u{}^i, \qquad k \in A, s \in S, i \in \Omega, \tag{4.6}$$

$$\sum_u p_{su}(\delta^i_{sk}) \xi_u{}^i = \xi_s{}^i - \mu^i_{sk}, \qquad k \in A, s \in S, i \in \Omega, \tag{4.7}$$

$$\sum_{i \in \Omega} \sum_{s \in S} \sum_{k \in A} (\lambda^i_{sk} + \mu^i_{sk}) D^i_{sk} = 0. \tag{4.8}$$

PROOF. $(\xi, \lambda, \mu, \omega, D)$ is a solution to (4.5)–(4.8) iff D generates $\delta \in \pi$ and δ attains

$$\xi_s{}^i + \omega_s{}^i = \max_{k \in A} \{r_s{}^i[\delta^i_{sk}] + \sum_u p_{su}(\delta^i_{sk}) \omega_u{}^i\}, \tag{4.9}$$

$$\xi_s{}^i = \max_{k \in A} \sum_u p_{su}(\delta^i_{sk}) \xi_u{}^i, \qquad s \in S, i \in \Omega, \tag{4.10}$$

which by [2], [6], [13] is necessary and sufficient for (2.1). □

REFERENCES

[1] BENIEST, W. (1963). Jeux stochastiques totalement cooperatifs arbitres. *Cahiers du Centre d'Etudes de Recherche Operationnelle* **5** 124–138.

[2] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[3] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[4] BLACKWELL, D. and FERGUSON, T. S. (1968). The big match. *Ann. Math. Statist.* **39** 159–163.

[5] CHARNES, A. and SCHROEDER, R. G. (1967). On some stochastic antisubmarine games. *Naval Res. Logist. Quart.* **14** 291–311.

[6] DENARDO, E. V. (1971). Markov renewal programming with small interest rates. *Ann. Math. Statist.* **42** 477–496.

[7] DENARDO, E. V. and FOX, B. L. (1967). Multichain Markov renewal programs. *SIAM J.* **16** 468–487.

[8] DERMAN, C. (1962). On sequential decisions and Markov Chains. *Man. Sci.* **9** 16–24.

[9] DERMAN, C. (1965). Markovian sequential control processes—denumerable state space. *J. Math. Anal. Appl.* **10** 295–302.

[10] EVERETT, H. (1957). Recursive games in *Contributions to the Theory of Games*, ed. M. Dresher, A. W. Tucker, and P. Wolfe. Princeton Univ. Press, 47–48.

[11] GILLETTE, D. (1957). Stochastic games with zero stop probabilities. In *Contributions to the Theory of Games, op. cit.* 179–188.

[12] HOFFMANN, A. J. and KARP, R. M. (1966). On non-terminating stochastic games. *Man. Sci.* **12** 359–370.

[13] HOWARD, R. (1960). *Dynamic programming and Markov processes*. Technology Press and Wiley, New York.

[14] LIGGETT, T. M. and LIPPMANN, S. A. (1968). Stochastic games with perfect information and time average payoff. Working Paper 142, Western Management Sci. Inst., Univ. of California, Los Angeles.

[15] MILLS, H. (1960). Equilibrium points in finite games. *SIAM J.* **8** 397–401.

[16] NASH, J. (1950). Equilibrium points in n-person games. *Proc. Nat. Acad. Sci. U.S.A.* **36** 48–49.

[17] NASH, J. (1951). Non-cooperative games. *Ann. of Math.* **54** 286–295.

[18] POLLATSCHEK, M. A. and AVI-ITZHAK, B. (1969). Algorithms for stochastic games. *Man. Sci.* **15** 399–415.

[19] RAPAPORT, A. and COLE, N. S. (1968). Experimental studies of interdependent mixed-motive games. *Behavioral Science* **13** 189–204.

[20] ROGERS, P. D. (1969). Nonzero-sum stochastic games. Report ORC 69–8, Operations Res. Center, Univ. of California, Berkeley.

[21] SCHWEITZER, P. J. (1968). Randomized gain-optimal policies for undiscounted Markov renewal programming. Unpublished manuscript, Institute for Defense Analyses, Arlington, Va.

[22] SHAPLEY, L. (1953). Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.* **39** 1095–1100.

[23] SOBEL, M. J. (1970). An algorithm for a game equilibrium point. Discussion Paper No. 7035, CORE, Univ. Catholique de Louvain, Belgium.

[24] VEINOTT, A. F. Jr. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.

[25] WOLF, G. and ZAHN, G. L. (1969). Exchange in games and communication. Unpublished manuscript, Dept. of Admin. Sci., Yale Univ.

[26] ZACHRISSON, L. E. (1964). Markov games in *Advances in Game Theory*, ed. M. Dresher, L. Shapley, and A. W. Tucker. Princeton Univ. Press, 211–253.