Conditions for equilibrium strategies in non-zero sum stochastic games

Citation for published version (APA):

Document status and date:
Published: 01/01/1980

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Conditions for equilibrium strategies in non-zero sum stochastic games.

by

J. Wessels

Eindhoven, November 1980

The Netherlands.
It is exhibited in this paper how one may obtain necessary and/or sufficient conditions for compound strategies to be equilibrium strategies in stochastic games. The method essentially consists of the construction of optimality conditions for a well-chosen set of one-person decision processes. The method is illustrated by the construction of necessary and sufficient conditions for equilibrium with a dynamic programming approach.

INTRODUCTION

In this paper we will consider the problem of how to construct necessary and/or sufficient conditions for a compound strategy to be an equilibrium strategy in a non-zero sum stochastic game. For technical simplicity, we will only consider discrete games, i.e. it is supposed that the state space and the action spaces for the players are countable, time is discrete and the number of players is finite. As equilibrium concept, we will consider the Nash-equilibrium concept and some of its variants.

The main technique for the construction of equilibrium conditions uses the fact that a compound equilibrium strategy should be built-up from parts which are in fact optimal strategies in some specific one-person decision problems. So, in general, it is possible to construct conditions for equilibriumnss from conditions for optimality in these one-person decision problems. In this paper we will first (section 2) exhibit the type of one-person decision problems which have to be studied in order to construct conditions for equilibriumnss in stochastic games. These one-person decision processes are essentially more complicated than the original games. Only if specially structured strategies are considered, these one-person decision processes have a structure which is more or less the same as the structure of the stochastic game. In section 3 this will be illustrated by constructing typical dynamic programming conditions.

1. The Model

The model we will consider in this paper, is a straightforward extension of Shap-
ley's model for zero-sum stochastic games [15] (see also Rieder [13] and van der Wal/Wessels [17]). A system can be observed at time points \( t = 0, 1, 2, \ldots \). The possible states of the system are \( 1, 2, \ldots \). After observation of the system at a point of time, the \( L \) players can each take an action from the set \( \{1, 2, \ldots \} \). If the system is in state \( i \) at time \( t \) and the players choose actions \( a^{(1)}, \ldots, a^{(L)} \) respectively, the result is that the probability of being in state \( j \) at time \( t+1 \) is equal to

\[
p(i, j; a) \geq 0, \text{ with } a = (a^{(1)}, \ldots, a^{(L)}) \text{ and } \sum_{j=1}^{\infty} p(i, j; a) = 1
\]

Moreover, player \( \ell \) earns an immediate reward of \( r_{\ell}(i; a) \). A compound strategy \( s \) consists of strategies for all players: \( s = (s^{(1)}, \ldots, s^{(L)}) \). A strategy \( s^{(\ell)} \) for player \( \ell \) consists of decision rules for player \( \ell \) for all timepoints \( t=0, 1, \ldots \):

\[
s^{(\ell)} = (s_0^{(\ell)}, s_1^{(\ell)}, \ldots) \]

A decision rule \( s_t^{(\ell)} \) for player \( \ell \) at time \( t \) designates the probabilities with which player \( \ell \) will choose the possible actions for each possible history until time \( t \):

\[
s_t^{(\ell)}(i_0, a_0, i_1, \ldots, a_{t-1}, i_t, a^{(\ell)}) \]

gives the probability for choosing decision \( a^{(\ell)} \) at time \( t \) by player \( \ell \) if the states \( i_0, \ldots, i_t \) and the compound actions \( a_0, \ldots, a_{t-1} \) have realized so far.

The history \( i_0, a_0, \ldots, a_{t-1}, i_t \) will be denoted by \( h_t \). In the obvious way one may show that a starting state \( i \) and a strategy determine a stochastic process \( (I_t, A_t) \), \( t = 0, 1, 2, \ldots \), where \( I_t \) denotes the state of the system at time \( t \) and \( A_t \) the compound action (throughout this paper, random variables will be denoted by capitals). The probability measure belonging to this process will be denoted by \( P_{i, s} \) and the expectation operator belonging to this probability measure will be denoted by \( E_{i, s} \). As criterion for player \( \ell \), we will consider the expected total reward for this player:

\[
v_{\ell}(i; s) \equiv E_{i, s} \left[ \sum_{t=0}^{\infty} r_{\ell}(I_t; A_t) \right].
\]

In order to guarantee existence and keep the analysis simple, we assume absolute convergence, i.e. finiteness of the same form with \( r_{\ell} \) replaced by \( |r_{\ell}| \).

Finally, the equilibrium concepts which will be considered are formulated briefly. The main concept is the concept Nash-equilibrium strategy. A strategy \( \sigma \) is said to be a Nash-equilibrium strategy if

\[
v_{\ell}(i; \sigma | s^{(\ell)}) \leq v_{\ell}(i; u), \text{ for all } \ell, i, s^{(\ell)},
\]

where \( \sigma | s^{(\ell)} \) denotes the strategy \( \sigma \) with \( \sigma^{(\ell)} \) replaced by \( s^{(\ell)} \).
So in fact $\sigma$ is a Nash-equilibrium strategy if $\sigma^{(t)}$ is a maximizer of $v_{\ell}(i;\sigma^{(t)})$. This leads to an alternative formulation in terms of a function $w_{\ell}$ defined by

$$w_{\ell}(i;\sigma) \equiv \sup_{s^{(t)}} v_{\ell}(i;\sigma | s^{(t)}).$$

Namely: $\sigma$ is a Nash-equilibrium strategy if

$$w_{\ell}(i;\sigma) = v_{\ell}(i;\sigma) \text{ for all } i,i.$$

The stronger equilibrium concepts will be formulated directly in this alternative form. Therefore we also need the expected total reward from time $t$ on, given some history until time $t$:

$$v(h_{t};s) = \mathbb{E}_{h_{t},s} \left( \sum_{t'=t}^{\infty} r_{h}^{(t)}(i_{t};a_{t}) \right),$$

where $\mathbb{E}_{h_{t},s}$ denotes the conditional expectation given that $h_{t}$ has materialized. If the history $h_{t}$ has probability 0 under strategy $s$ and starting state $i_{0}$, then an appropriate definition of $v_{\ell}(h_{t};s)$ is the expectation of $\sum_{t'=t}^{\infty} r_{h}^{(t)}(i_{t};a_{t})$ for the process starting at time $t$ in state $i_{t}$ with the compound strategy $s$ applied as if the path $h_{t}$ has materialized. In this way $\mathbb{E}_{h_{t},s}$ may be considered to be well-defined for all $h_{t},s$ and it coincides with the conditional expectation given $h_{t}$ if this path has a positive probability under strategy $s$.

Also the domain of the function $w_{\ell}$ can be extended:

$$w_{\ell}(h_{t};\sigma) \equiv \sup_{s^{(t)}} v_{\ell}(h_{t};\sigma | s^{(t)}).$$

Note also that $w_{\ell}(h_{t};\sigma)$ and $v_{\ell}(h_{t};\sigma)$ do not depend on $\sigma_{0},\sigma_{1},\ldots,\sigma_{t-1}$. Moreover, $w_{\ell}$ does not depend on $\sigma_{t}^{(t)},\sigma_{t+1}^{(t)}$.

For a Nash-equilibrium strategy $\sigma$ one has

$$(*) \quad w_{\ell}(h_{t};\sigma) = v_{\ell}(h_{t};\sigma)$$

for all $t,\ell$ and those $h_{t}=(i_{0},a_{0},\ldots,i_{t})$ which have a positive probability under strategy $\sigma$ with starting state $i_{0}$.

In fact this is an equivalent formulation of the concept Nash-equilibrium strategy and in the sequel it will be used as a definition. Also the stronger equilibrium concepts which will be considered, can be formulated in terms of $w_{\ell}$. These stronger concepts only differ from Nash's concept in the sense that $(*)$ is required for a larger set of paths $h_{t}$. 
A strategy \( \sigma \) is said to be semi-subgame perfect (a concept introduced by Couwenbergh in [1] under a different name; here we use the name given by Groenewegen in [4]), if (*) holds for all \( t, \ell \) and those \( h_t \) which have a positive probability under \( \sigma|s(\ell) \) for some \( s(\ell) \) with starting state \( i_0 \).

A strategy \( \sigma \) is said to be tail-optimal (a concept introduced by Groenewegen in [4]), if (*) holds for all \( t, \ell \) and those \( h_t \) which have a positive probability under \( \sigma|s(m) \) for some \( s(m) \) (any \( m \)) with starting state \( i_0 \).

A strategy \( \sigma \) is said to be sub-game perfect (a concept introduced by Selten, see [14]), if (*) holds for all \( t, \ell \) and those \( h_t \) which have a positive probability under some strategy \( s \) with starting state \( i_0 \).

2. Reduction to one-person decision problems

As has been remarked already in the preceding section, a Nash-equilibrium strategy \( \sigma = (\sigma^{(1)}, \ldots, \sigma^{(L)}) \) has the property that any of its constituents \( \sigma^{(k)} \) is a maximizing strategy in the one-person decision problem with criterion function \( v_k(i; \sigma|s^{(k)}) \). Conversely, if \( \sigma \) has this property, then \( \sigma \) is a Nash-equilibrium strategy.

So, if one is interested in the construction of necessary and/or sufficient conditions for a compound strategy to be a Nash-equilibrium strategy, then this problem may be solved by giving optimality conditions for the relevant one-person decision problems.

Hence the problem of constructing conditions for strategy \( \sigma \) in the stochastic game has been reduced to constructing conditions for the strategies \( \sigma^{(1)}, \ldots, \sigma^{(L)} \) in one-person decision problems with criteria \( v_1(i; \sigma|s^{(1)}), \ldots, v_L(i; \sigma|s^{(L)}) \) respectively. The criterion function \( v_k(i; \sigma|s^{(k)}) \) belongs to the one-person decision process which is obtained from the game by fixing the strategies \( \sigma^{(1)}, \ldots, \sigma^{(k-1)}, \sigma^{(k+1)}, \sigma^{(L)} \) and only varying \( s^{(k)} \). Regrettably, this decision process does not have a simple structure if \( \sigma \) is arbitrary. Only if \( \sigma \) is a Markov strategy (i.e. the \( \sigma^{(m)} \) don't really depend on \( h_t \), but only on \( i_t \), then the relevant one-person decision process is a (nonstationary) Markov decision process. If \( \sigma \) is even stationary (i.e. \( \sigma^{(m)} \) does not depend on \( t \) either), then we obtain a stationary Markov decision process.

For stationary and nonstationary Markov decision processes, simple and useful optimality conditions can be given. For more general decision processes optimality conditions may be elegant, however, they are not easy to operationalize.

Regrettably, it is not sensible to restrict attention to Markov strategies, as one might think at first sight. It is well-known already that such a restriction would possibly exclude some very interesting strategies. For several examples and more
references on the subject the reader is referred to the paper by van Damme [2] in this volume.

Hence, since inclusion of history-dependent strategies is essential, we are faced with the fact that the one-person decision processes which have to be analyzed in order to obtain equilibrium conditions for stochastic games, have an essentially more complicated structure than the stochastic games. This also remains true if one considers specific types of stochastic games, such as difference games.

Let us now study more precisely which type of one-person decision process has to be analyzed. If player \(i\) has decided on decision \(d\) at time \(t\), after history \(h_t\) has been realized and other players use strategy \(s\), then the one-stage reward for player \(i\) is determined by the outcomes of random experiments for \(m \neq i\) with probability distributions \(\sigma^{(m)}(h_t, \cdot)\). The same holds for the transition probabilities.

By introducing these lottery outcomes as a supplementary variable \(Y_t(Y_t \text{ is } A_t \text{ with } A_t^{(i)} \text{ discarded}), we may describe the one-person decision process by \((I_t, Y_t, D_t)\). The player \(i\) may base his decisions at time \(t\) on \(I_t(t = 0, \ldots, t), Y_t(t = 0, \ldots, t-1)\). Hence an appropriate reformulation would be the following.

Take \(X_t = (I_t, Y_{t-1})\) as state of the one-person decision process. Then the immediate reward at time \(t\) (for this one player) depends on \(I_t, Y_t, D_t\), hence on \(X_t, D_t\) and \(X_{t+1}\). The transition probabilities depend on \(X_0, \ldots, X_t\) and \(D_0, \ldots, D_t\). So we have to study optimality conditions for one-person decision processes with countable state space, countable action space, immediate reward at time \(t\) equal to \(p(x, d, x')\) if \(x\) is the state at time \(t\), \(d\) the decision and \(x'\) the resulting state at time \(t\). The transition probabilities \(q(x_0, d_0, \ldots, d_{t-1}, x_t; x, d)\) give the probability for reaching state \(x\) at time \(t+1\), given the history \(x_0, d_0, \ldots, x_t\) and the current decision \(d\).

3. A set of equilibrium conditions

In the preceding section, it has been shown that the problem of finding sufficient and/or necessary conditions for a compound strategy in a stochastic game to be a Nash-equilibrium strategy, can be reduced to the construction of optimality conditions for a one-person decision process. For that type of decision process one may proceed in different ways when seeking optimality conditions. If no further analytic structure has been given (difference games - for instance - do have a very well exploitable analytic structure; convexity in some of the parameters also provides a very well exploitable analytic structure via Rockafellar's convex analysis, see Klein Haneveld [11]), then there are essentially two well-known ways. The first one uses a linear programming approach (for linear programming treatments of rather general types of one-person decision processes see Heilmann [9] and Klein Haneveld
the second one uses a dynamic programming approach. As an illustration we will work out conditions according to a dynamic programming approach. In doing so we follow a line of work on optimality conditions for dynamic decision problems which starts with Shapley and Bellman. In this line Dubins and Savage [3] and Sudderth [16] introduced some essential notions, which have been transmitted to Markov decision processes by Hordijk [10] and brought into a very general setting by Groenewegen [4], see also [8].

Before formulating optimality conditions, we need some notations and definitions for the one-person decision process. These notations and definitions are analogous to the notations and definitions for the stochastic game and will therefore be treated only briefly.

Strategies (history dependent) for the one-person decision process are denoted by \( \pi, \pi' \) etc. A strategy \( \pi \) is a sequence \( (\pi_0, \pi_1, \ldots) \) of decision rules in which each \( \pi_t \) designates the probabilities with which the actions will be chosen for each possible history until time \( t : \pi_t(g_t; d) \), with \( g_t = (x_0, d_0, \ldots, x_{t-1}, d_{t-1}, x_t) \), gives the probability of selecting action \( d \) if history \( g_t \) has materialized. \( P_{x,\pi} \) and \( E_{x,\pi} \) denote the probability measure and expectation operator belonging to the stochastic process \( (X_t, D_t), t = 0, 1, \ldots \) for given starting state \( x \) and strategy \( \pi \). \( X_t \) and \( D_t \) are random variables denoting the state at time \( t \) and the action at time \( t \). Also the conditional expectation operator \( E_{g_t,\pi} \) may be introduced as in the stochastic game and extended to \( g_t \) with probability 0. Important quantities are again the total expected rewards from time \( t \) on, given some history \( g_t \):

\[
V(g_t; \pi) = E_{g_t,\pi} \left( \sum_{t=0}^{\infty} \rho(X_t, D_t, X_{t+1}) \right)
\]

A strategy \( \pi' \) is optimal if

\[
V(x; \pi') = \sup_{\pi} V(x; \pi) \text{ for all } x.
\]

This optimal value will be called \( w(x) \) henceforth. For an optimal strategy \( \pi' \), we also have

\[
V(g_t; \pi') = \sup_{\pi} V(g_t; \pi) \text{ for all } t \text{ and } \pi
\]

those \( g_t \) which have a positive probability under strategy \( \pi' \) with starting state \( x_0 \). As for the games we extend the domain of definition for the function \( w \):

\[
w(g_t) = \sup_{\pi} V(g_t; \pi).
\]
Now we can formulate optimality conditions.

**Theorem 1**: A strategy \( \pi \) in the one-person decision process is optimal if and only if it is conserving and equalizing, i.e.

a. (conserving) \( \mathbb{E}(G_t) = \mathbb{E}_{G_t, \pi} \left[ \sum_{t=1}^{\infty} \rho(X_t, D_t, X_{t+1}) \right] \)

when \( G_t \) denotes a random path.

b. (equalizing) \( \lim_{t \to \infty} \mathbb{E}_{x_0, \pi} \left[ \mathbb{E}(G_t) \right] = 0 \) for all \( x_0 \).

**Proof**: Both parts of the proof are very simple and can be based on the proof in [5] or [8]. Suppose \( \pi \) is conserving and equalizing. Then

\[
\mathbb{E}(x_0) = \mathbb{E}_{x_0, \pi} \left[ \sum_{t=1}^{\infty} \rho(X_t, D_t, X_{t+1}) \right] + \mathbb{E}_{x_0, \pi} \left[ \mathbb{E}(G_t) \right]
\]

Hence \( \pi \) is optimal.

For the proof of the necessity of the conditions the essential point is that for an optimal strategy (in fact for any strategy)

\[
\lim_{t \to \infty} \mathbb{E}_{x_0, \pi} \left[ \mathbb{E}(G_t) \right] = 0.
\]

Based upon this set of necessary and sufficient conditions for optimality in one-person decision processes, we obtain necessary and sufficient conditions for some strategy \( \sigma \) to be a Nash-equilibrium strategy in a stochastic game. Similarly as in theorem 1, we denote by \( H_t \) the random path \( (I_0, A_0, \ldots, I_t) \).

**Theorem 2**: The compound strategy \( \sigma \) is a Nash-equilibrium strategy if and only if it is conserving and equalizing i.e.

a. (conserving) \( \mathbb{E}(H_t) = \mathbb{E}_{H_t, \sigma} \left[ \sum_{t=1}^{\infty} \rho(I_t, A_t) \right] \)
Theorem 2 now simplifies to

**Corollary 2**: The compound Markov strategy \( \sigma \) is a Nash-equilibrium strategy if and only if

\[
\text{a. } w_k(t; I_t; \sigma) = E_{I_t, \sigma} \left( r_k(I_t; A_t) + w_k(t+1; I_{t+1}; \sigma) \right) \\
\text{b. } \lim_{t \to \infty} E_{I_t, \sigma} \left( w_k(t; I_t; \sigma) \right) = 0 \text{ for all } i.
\]

The analogous simplification can be executed for theorem 3, if \( \sigma \) is a Markov strategy.

For Markov strategies these conditions may be reformulated in matrix form. Namely, denote by \( P(\sigma) \) the matrix with \((i, j)\)-entry the transition probability to state \( j \).
from state $i$ if compound decision rule $\sigma_t$ is used by the players. $r_k(\sigma_t)$ denotes in the same way the vector of expected one-stage rewards for the different states for player $k$ if decision rule $\sigma_t$ is used. Then corollary 2 can be rewritten as

**Corollary 2** : The compound Markov strategy $\sigma$ is a Nash-equilibrium strategy if and only if

a. $w_k(t;\sigma) = r_k(\sigma_t) + P(\sigma_t) w_k(t+1;\sigma)$ for those components $j$ for which
   
   $P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot P(\sigma_t)(i,j) > 0$ for some $i$ (for all $k,t$);

b. $\lim_{t \to \infty} P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot w_k(t;\sigma) = 0$.

Note that $w_k(t;\sigma)$ is the vector with components $w_k(t;i;\sigma)$.

For the stronger equilibrium concepts we obtain in this way

**Corollary 3** : The compound Markov strategy $\sigma$ is semi-sub game perfect or tail-optimal or sub-game perfect respectively if and only if the conditions a and b below hold for all $k,t$ and for those components $j$ for which

- $[P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot P(\sigma_t)](i,j) > 0$ for some $i,s$ (Markov strategy);
- $[P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot P(\sigma_t)](i,j) > 0$ for some $i,m,s$ or
- $[P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot P(\sigma_t)](i,j) > 0$ for some $i,m,s$ respectively

a. $w_k(t;\sigma) = r_k(\sigma_t) + P(\sigma_t) w_k(t+1;\sigma)$;

b. $\lim_{t \to \infty} P(\sigma_0) \cdots P(\sigma_{t-1}) \cdot w_k(t;\sigma) = 0$.

For stationary strategies the conditions simplify further. In fact the four equilibrium concepts coincide for stationary strategies as can also be seen from the conditions. $w_k$ no longer depends on $t$.

**Corollary 2** : The compound stationary Markov strategy $\sigma$ is an equilibrium strategy in all 4 senses as introduced if and only if for all $k$

a. $w_k(\sigma) = r_k(\sigma_0) + P(\sigma_0) w_k(\sigma)$;

b. $\lim_{t \to \infty} p^t(\sigma_0) \cdot w_k(\sigma) = 0$. 
For Markov strategies the conditions given here are the same as the conditions derived by Couwenbergh [1] and Groenewegen [4] using the analogy of dynamic games with Markov decision processes. That approach, however, only works for Markov strategies. The conditions for general strategies of theorem 2 and 3 can also be obtained by specification from theorems in Groenewegen [5] and Groenewegen/Wessels [7]. This shows that our method of constructing conditions for dynamic games is effective and might be useful to transfer other types of conditions from one-person decision processes to dynamic games.

REFERENCES

[1] Couwenbergh, H.A.M., Characterization of strong (Nash) equilibrium points in Markov games, Memorandum COSOR 77-09 (April 1977) Eindhoven University of Technology (Dept. of Math.).


