Mitri Kitti

Subgame Perfect Equilibria in Discounted Stochastic Games

Aboa Centre for Economics
Discussion paper No. 87
Turku 2013

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku 2013

Mitri Kitti

Subgame Perfect Equilibria in Discounted Stochastic Games

Aboa Centre for Economics
Discussion paper No. 87
December 2013

ABSTRACT

This paper considers policies and payoffs corresponding to subgame perfect equilibrium strategies in discounted stochastic games with finitely many states. It is shown that a policy is induced by an equilibrium strategy if and only if it can be supported with the threat of reverting to the induced policy that gives the least equilibrium payoff for the deviator. It follows that the correspondence of subgame perfect equilibrium payoffs is the largest fixed-point of a correspondence-valued operator defined by the players' incentive compatibility conditions. Moreover, the fixed-point iteration converges to the equilibrium payoff correspondence.

JEL Classification: C73

Keywords: subgame perfect equilibrium, stochastic game, payoff correspondence, fixed-point equation, induced equilibrium policy

Contact information

Mitri Kitti
Department of Economics
University of Turku
FI-20014, Finland
Email: mitri.kitti (at) utu.fi

1. Introduction

Subgame perfection is the most important solution concept for stochastic games. Among the major challenges in analyzing such equilibria is their multiplicity. This paper provides results to tackle this issue. Most importantly, I derive a fixed-point equation for the equilibrium payoff correspondence, i.e., the set-valued mapping from initial states to players' equilibrium payoffs. It is assumed that there are finitely many states, players have perfect monitoring, and they use discounted payoffs as the evaluation criterion of strategies.

The characterization of the equilibrium payoff correspondence builds upon *extremal penal codes*, which are composed of *induced equilibrium policies* that lead to the players' smallest equilibrium payoffs. A policy is called an induced equilibrium policy if it describes play of the game when the players' follow subgame perfect equilibrium strategies. The extremal penal codes, on the other hand, represent the most severe threats for deviating players.

When deviations from an induced equilibrium policy lead to punishments determined by the extremal penal code, then no player has an incentive to deviate. More specifically, a necessary and sufficient condition for a policy to be an equilibrium outcome is that the *simple strategy* in which the policy is supported with an extremal penal code is subgame perfect. Simple strategies and penal codes were originally introduced by Abreu [1, 2] for repeated games. Kitti [16, 17] has recently defined stationary and Markov penal codes in dynamic games with perfect monitoring, i.e., in games where players observe each others' actions and the random disturbances affecting the payoffs and state transitions.

If we are given a profile of players' actions at a given state together with a continuation value function that determines the player's future payoffs, we can test whether this pair of action profiles and continuation values could represent an equilibrium outcome. Namely, if all players prefer the action profile and the given payoff function for deviating and receiving their least equilibrium payoffs, then we could expect the pair to represent equilibrium behavior. By utilizing this idea of incentive compatibility we can define an operator for payoff correspondences that gives us the discounted payoffs corresponding to those action profiles and continuation functions that players prefer to deviations. It is shown that the equilibrium payoff correspondence is the largest fixed-point of this operator with respect to set-inclusion.

The result for the equilibrium payoff correspondence is analogous to those obtained for repeated games, see [3, 4] for the case of imperfect monitoring, and [13] for the case of perfect monitoring. In the dynamic game setup the set-valued approach has been used in analyzing a deterministic dynamic game of greenhouse gas emissions [14]. Kitti [17] has recently provided a

fixed-point characterization for equilibrium payoffs of stochastic games that correspond to conditional Markov strategies.

The fixed-point result for the equilibrium payoff correspondence can be used for computational purposes. In this paper it is shown that the equilibrium payoff correspondence can be found by using a set-valued fixed-point iteration. For repeated games related iterations have been studied in [12] and [15], and for correlated strategies in deterministic dynamic games in [23]. In the framework of stochastic games, Mertens and Parthasarathy [19, 20] show the existence of equilibria by iterating a particular payoff correspondence. However, their iteration is different from the one analyzed in this paper.

The paper is structured as follows. The model and the main assumptions are presented in Section 2. Simple strategies and extremal penal codes are introduced in Section 3. The characterization of equilibrium payoff correspondence is presented in Section 4. Convergence of the fixed-point iteration is analyzed in Section 5. In this paper the main attention will be on pure strategies. Extension to the case of behavior strategies is, however, discussed in Section 6 for games with finitely many actions and states.

2. The Model

There are n players indexed with $i \in I = \{1, \ldots, n\}$. The players' available actions at state $x \in X$ are $Y_i(x)$, $i \in I$. It is assumed that X is a finite set. The correspondence of feasible action profiles is denoted as

$$Y(x) = Y_1(x) \times \cdots \times Y_n(x), \text{ and } Y = \cup_{x \in X} Y(x).$$

The state evolves according to a dynamic system

$$x^{k+1} = f\left(y^k, x^k, z^k\right), \ k = 0, 1, \ldots,$$

where f is a function from $Y \times X \times Z$ into the set of states X. Here Z is a finite set of random disturbances affecting the evolution of the state and the players' payoffs. The players choose their actions simultaneously after which the disturbance is realized and the state transition takes place. The random disturbances are identically distributed with probabilities $\text{Prob}(z|y, x)$, $z \in Z$.

The per-period payoffs are determined by functions $u_i : Y \times X \times Z \mapsto \mathbb{R}$, $i \in I$. It is assumed that the payoff sets

$$\{(u_1(y, x, z), \ldots, u_n(y, x, z)) \in \mathbb{R}^n : y \in Y(x)\}$$

are compact for all $x \in X$ and $z \in Z$. The players observe perfectly each others' actions and the disturbances, and hence the states. Note that, the

states are fully determined from the dynamic system governing the state transitions when the initial state, and the past actions and disturbances are known.

The duration of the game is infinite. A history in stage k is denoted as h^k and it is defined recursively as follows: $h^0 = x^0 \in X$, and for $k \geq 1$ the history is $h^k = h^{k-1} \cup \{y^{k-1}, z^{k-1}\}$. The set of all possible histories in stage k for any initial state is denoted as H^k and the set of all histories is H. Moreover, $x(h^k)$ is the state that has been reached after the history $h^k \in H^k$.

A strategy for player i is a sequence of functions $(\sigma_i^0, \sigma_i^1, \ldots)$ where σ_i^k maps any $h^k \in H^k$ into $Y_i(x(h^k))$. A strategy profile is denoted as $\sigma = (\sigma_i, \ldots, \sigma_n)$ and, as usual, σ_{-i} is the collection of strategies of other players than player i. In this paper the attention is on pure strategies. However, all the concepts generalize directly to behavior strategies, when strategies are defined as mappings into distributions over $Y_i(x)$. This will be discussed in Section 6.

Each player i chooses a strategy σ_i which maximize the expected discounted payoff

$$\lim_{m\to\infty} \mathbb{E}\left[\sum_{k=0}^{m} \delta_i^k u_i\left((\sigma_i^k(h^k), \sigma_{-i}^k(h^k)), x(h^k), z^k\right)\right],$$

where $\delta_i \in (0,1)$ is the discount factor and the expectation is over disturbance sequences $z^0, z^1, \ldots$. Note that the above objective function is well-defined and finite because X and Z are finite, and payoff sets are compact.

When the players follow σ, the expected payoff for player i after k-period history h^k is

$$U_i\left(\sigma, h^k\right) = \lim_{m\to\infty} \mathbb{E}\left[\sum_{j=0}^{m} \delta_i^j u_i\left(\sigma^{k+j}(h^{k+j}), x(h^{k+j}), z^{k+j}\right)\right]. \tag{1}$$

Subgame perfection is defined in the usual manner.

Definition 1. Strategy profile σ is a subgame perfect equilibrium (SPE) if for all $i \in I$ it holds that

$$U_i\left(\sigma, h\right) \geq U_i\left((\sigma_i', \sigma_{-i}), h\right)$$

for all strategies σ_i' of player i, and histories $h \in H$.

When players follow a given strategy σ, the expected future payoffs implied by the strategy at any stage for an action profile $y \in Y(x(h^k))$, are called continuation payoffs of the strategy. Hence, if players take an action

profile y at stage k for the initial history $h^k \in H^k$, the payoff for player i is $\mathbb{E}_z[u_i(y, x(h^k), z) + \delta_i v_i(f(y, x, z))]$, where v_i is player i's continuation payoff function, i.e., $v_i(x)$ is the payoff corresponding to $\sigma^{k+1}(h^k \cup \{y, z\})$ for initial state x. The vector valued functions $v = (v_1, \ldots, v_n)$ will be also called continuation payoff functions. It is worth observing that the continuation payoffs corresponding to a strategy profile are defined only for states that can be reached. For instance, if there is an absorbing state $x^0 \in X$, the continuation payoff function of σ after h^k with $x(h^k) = x^0$ need not be defined for other states than x^0.

In the rest of the paper, Σ denotes the set of subgame perfect equilibrium strategies. The correspondence V will denote the subgame perfect equilibrium payoffs for all initial states. To be specific, the value of the equilibrium payoff correspondence V at $x \in X$ is

$$V(x) = \{U(\sigma, x) \in \mathbb{R}^n : \sigma \text{ is an SPE when } x^0 = x\}.$$

3. Simple Strategies

In this section the purpose is to provide necessary and sufficient conditions for behavior induced by equilibrium strategies. It will be observed that equilibrium behavior is determined by policies, which will be called induced equilibrium policies. The main result of this section is that a policy is an induced equilibrium policy if and only if it is supported by the threat of reverting to the induced equilibrium policy which leads to the smallest equilibrium payoff for the deviator. This result builds upon the concepts of simple strategy and penal code, which will be defined in the spirit of Abreu [1, 2].

3.1. Induced Equilibrium Policies

As long as players follow their strategies for a given initial history, they are said to follow a policy induced by the strategy and the initial history. For example, if the initial state is $x^0 \in X$, the strategy profile gives the action profile $y^0 = \sigma^0(x^0)$ in the first period. In the next period the action profile $\sigma^1(x^0, y^0, z^0)$ is played according to the strategy profile σ, and so on for σ^k, $k \geq 2$. We can observe that as long as all players follow σ the actions will depend on the past history of disturbances, the initial state, and actions taken in the first period.

Let us now assume that a history $h \in H^j$ has realized, and during this history some of the players has made a unilateral deviation. Then according to σ the players take actions $y^j = \sigma^j(h)$ in period j, and after that the actions are $\sigma^{j+1}(h, y^j, z^j)$, $\sigma^{j+2}(h, \sigma^{j+1}(h, y^j, z^j), z^{j+1}), \ldots$. As a result, the

equilibrium play after the initial history $h \in H^j$ is a sequence of the form $y^0, \nu^1(z^0; h, y^0), \nu^2(z^0, z^1; h, y^0), \ldots$. Observe that from period j onwards the initial history h, which determines the first action profile y^0, is fixed. The future actions can depend on the realized h, but when viewed from period j onwards, the action profiles in any period $j+k$, $k \geq 1$, are just functions from disturbance sequences from periods j to $j+k-1$ into feasible actions, i.e., $\nu^k(\cdot; h, y^0)$ maps sequences of Z^k into feasible action profiles.

A sequence $y^0, \nu^1, \nu^2, \ldots$ corresponding to the strategy profile σ and a given initial history h is called an induced policy of σ after $h \in H^j$, $j \in \{0, 1, \ldots\}$. As argued above, when viewed from period j onwards and assuming that no-one deviates from σ, the players' actions will only depend on the stage of the game and the sequence of past disturbances from period j onwards. The dependence on the history of players' actions becomes relevant only when some of them deviates from the ongoing policy. After a deviation the play of the game will follow an induced policy chosen according to the past history of actions and disturbances.

In the following a policy will refer to a sequence $\mu^0, \mu^1, \mu^2, \ldots$, where μ^j maps elements of $X \times Z^j$ into feasible action profiles. Observe that an induced policy of σ for h , as described above, i.e., a sequence of the form $y^0, \nu^1, \nu^2, \ldots$, is not a policy of this kind. However, it will be more convenient to use the notion of policy in the common sense of dynamic programming literature, see, e.g., [10].[1] The set of policies will be denoted as Π. Let us now define the induced equilibrium policies such that they become policies in the above sense.

Definition 2. A policy $\pi \in \Pi$ is an induced equilibrium policy if for any $x \in X$ there are $\sigma \in \Sigma$, $k \geq 0$, and $h \in H^k$ with $x(h) = x$, such that the induced policy of σ for h gives the same action profiles as π after all disturbance sequences.

In practice, an induced equilibrium policy is obtained by choosing a different induced policy of the equilibrium strategy σ for h corresponding to any initial state. For instance, let us assume that there are two states, x^1 and x^2, and σ induces $y^1, \nu^1(z^0; h^1, y^1), \ldots$ after h^1 with $x(h^1) = x^1$, and $y^2, \nu^1(z^0; h^2, y^2), \ldots$ after h^2 with $x(h^2) = x^2$. Then the sequence $\mu^0, \mu^1, \ldots$ with $\mu^0(x^i) = y^i$ and $\mu^k(x^i, \cdot) = \nu^k(\cdot; h^i, y^i)$, $i = 1, 2$, $k \geq 1$, is an induced equilibrium policy.

[1] Usually, a policy is a sequence of mappings from histories of states into feasible actions. Except for the initial state, the history of states is replaced with the history of disturbances.

3.2. Equilibria in Simple Strategies

Before defining simple strategies we need some notations. The length k histories obtained when π is followed is denoted as $H^k(\pi)$. The set $H^k(\pi)(x)$ denotes the histories obtained by following π beginning from the initial state $x \in X$. The set $\text{ind}(\Sigma) \subseteq \Pi$ contains the induced equilibrium policies.

A simple strategy is composed of a set of a policy that the players follow until some of them makes a unilateral deviation, and policies that are followed after such deviations. After deviations players start to follow new policies corresponding to the deviator. Simultaneous deviations can be ignored as irrelevant because we are dealing with non-cooperative equilibria. The simple strategies of repeated games [1, 2] and the stationary and Markovian simple strategies [16, 17] are specific cases of the general simple strategies considered in this paper. In repeated games policies are simply paths of action profiles. Equilibrium paths of such games have been recently analyzed by Berg and Kitti [8, 7, 9] who utilize extremal punishments which will also play an important role in this paper.

Definition 3. A strategy profile is simple if

1. there is an initial policy $\pi^0 \in \Pi$ that the players follow until some of them deviates unilaterally,
2. when player i is the last who has unilaterally deviated from the ongoing policy, the players start to follow the policy $\pi^i \in \Pi$. Any subsequent unilateral deviations lead to π^i corresponding to the deviator $i \in I$

The composition of policies $p = \{\pi^1, \ldots, \pi^n\}$ is called a penal code. The simple strategy composed of a policy π^0 and a penal code p is denoted as $\sigma(\pi^0, p)$.

In the following $v(\pi)(x)$ denotes the vector of players' expected payoffs when they start to follow the policy $\pi \in \Pi$ from the initial state x. We shall be interested in the penal codes which lead to players' smallest equilibrium payoffs. These payoffs will be denoted as $v_i^-(x; V) = \inf\{v_i : v \in V(x)\}$, $i \in I$, $x \in X$. The set of players' continuation payoff functions that are obtained by selecting the components from $V(x)$ for all $x \in X$, is denoted as $F(V)$. Moreover, $F_i(V)$ denotes the corresponding payoffs for player i. In particular, it will be shown that $v_i^-(\cdot; V) \in F_i(V)$, $i \in I$.

Definition 4. The penal code $p = \{\pi^1, \ldots, \pi^n\}$ is an equilibrium if $\sigma(\pi^i, p) \in \Sigma$ for all $i \in I$. The penal code is extremal if it is an equilibrium and $v_i(\pi^i)(x) = v_i^-(x; V)$ for all $x \in X$.

As shown below, there are extremal penal codes. The result follows from the one-shot deviation principle [11], which says that it is optimal to follow a given strategy if and only if there are no profitable one-shot deviations from it at any stage. The reason why extremal penal codes are important is that they support all equilibrium policies. More specifically, a necessary and sufficient condition for $\pi \in \Pi$ to be an induced equilibrium policy is that the simple strategy $\sigma(\pi, p^*)$ is an SPE.

Proposition 1. *The following results hold.*

1. *An extremal penal code p^* exists when $V(x) \neq \emptyset$ for all $x \in X$.*
2. *$\pi \in ind(\Sigma)$ if and only if $\sigma(\pi, p^*) \in \Sigma$.*
3. *V is compact valued.*

Proof. Let us begin by showing that extremal penal code exists when there are subgame perfect equilibria. Let us pick a sequence $\{v^j\}_j \in F(V)$ converging to v with $v_i(x) = v_i^-(x; V)$ for all $x \in X$ and some $i \in I$. Note that the finiteness of X implies that $F(V)$ is a finite dimensional set, and hence the convergence can be regarded in the usual Euclidean metric. First, for any j there is π^j such that $v(\pi^j)(x) = v^j(x)$, $x \in X$. Moreover, it holds that

$$v_i^j(x) = \sum_{k=0}^{\infty} \delta_i^k \sum_{h^k \in H^k(\pi^j)(x)} \bar{u}_i\left(\pi^{k,j}\left(h^k\right), x^k\right) \mathrm{Prob}\left(h^k | \pi, x^0 = x\right),$$

where $\bar{u}_i(y, x) = \sum_{z \in Z} \mathrm{Prob}(z|y, x) u_i(y, x, z)$, and $\pi^{k,j}(h^k)$ is the action profile prescribed by the policy π^j in stage k after history $h^k \in H^k$. Because $H^k(\pi^j)(x)$ is finite for any $k \geq 0$, $x \in X$, and payoff sets are compact, we can use the diagonalization argument to construct a convergent subsequence such that $\bar{u}_i(\pi^{k,j}(h^k), x^k)$ and $\mathrm{Prob}(h^k|\pi, x^0 = x)$ converge for all $i \in I$, $x \in X$. It follows that there is also a limit policy $\pi^i \in \Pi$ yielding the payoffs v_i. Repeating the argument for all players we obtain π^i, $i \in I$.

Because v^j in the above deduction corresponds to an SPE strategy, one-shot deviation principle implies that none of the players wants to make any deviations from π^j at any stage given the continuation payoffs of the strategy. The latter are at least $v_i^-(x; V)$, $x \in X$. Hence there are no profitable one-shot deviations from π^j when the punishment payoffs for deviations are given by $v_i^-(\cdot; V)$, $i \in I$. Consequently, there are no profitable one shot deviations from π^i, $i \in I$, obtained in the limit. Hence, the penal code composed of $\pi^1, \ldots, \pi^n$ is an equilibrium, i.e., an extremal penal code exists.

The second result follows directly from the one-shot deviation principle and the fact that there are extremal penal codes. Namely, if we pick an induced equilibrium policy π, there are no profitable one shot deviations

from it at any stage given the players' continuation payoffs corresponding to the deviations. The latter are at least $v_i^-(\cdot;V)$, $i \in I$. Hence, there cannot be profitable one shot deviations when the players' continuation payoffs of the strategy are replaced with $v_i^-(\cdot;V)$, $i \in I$. Because $\sigma(\pi,p)$ is an SPE if and only if there are no profitable one-shot deviations at any stage, we get the result.

Let us next consider the last part of the proposition: the compactness of $V(x)$ for all $x \in X$. Similarly as for $v_i^-(\cdot;V)$, $i \in I$, we can construct a policy $\pi \in \Pi$ giving any limit payoff of convergent sequences of $F(V)$. Again, there are no profitable one shot deviations from these policies when the punishment payoffs are v_i^-. Hence, we get $\sigma(\pi,p^*) \in \Sigma$ and the third result follows. □

3.3. Example: Two-State Prisoners' Dilemma

This example presents a penal code in a dynamic game where the stage games are prisoners' dilemmas. A deterministic variation of the game has been discussed by Kitti [16]. There are two players and two states x_1 and x_2, $W = X$, and

$$Y_1(x_1) = Y_1(x_2) = \{a_1, a_2\},\ Y_2(x_1) = Y_2(x_2) = \{b_1, b_2\}.$$

The state transition is $f(y,x,w) = w$. At state x_1 the transition probabilities are $\text{Prob}(w = x_1|(a_i,b_j),x_1) = 1$ when either $i \neq 2$ or $j \neq 2$, and

$$\text{Prob}\,(w = x_1|(a_2,b_2),x_1) = \text{Prob}\,(w = x_2|(a_2,b_2),x_1) = 1/2.$$

At state x_2 the probabilities are $\text{Prob}(w = x_1|(a_i,b_j),x_2) = 1/2$ when either $i \neq 2$ or $j \neq 2$, and $\text{Prob}(w = x_2|(a_2,b_2),x_2) = 1$. The payoffs are as below where $*$ signifies the action profiles from which transition to the other state is possible. Observe that the stage games are variations of the prisoners' dilemma game.

x_1	b_1	b_2
a_1	$(4,4)$	$(0,5)$
a_2	$(5,0)$	$(1,1)^*$

x_2	b_1	b_2
a_1	$(0,0)^*$	$(-4,1)^*$
a_2	$(1,-4)^*$	$(-3,-3)$

The penal code p^* is composed of the following policies π^i, for $i \in I$. Corresponding to the first player, the row player, π^1 is the policy profile in which (a_1,b_2) is played in every period in both states, and for the second player π^2 is the policy profile in which (a_2,b_1) is played in both states. It can be seen that there is no incentive to deviate from $\sigma(\pi^i,p^*)$ when $\delta \geq 2/5$, i.e.,

p^* is an equilibrium penal code. These policies form an extremal penal code because they lead to min-max payoffs, i.e., players' security levels which are

$$v_i(\pi^i)(x) = \begin{cases} 0, & x = x_1 \\ -8/(2-\delta), & x = x_2. \end{cases}$$

If we take a common discount factor $\delta < 2/5$, then the extremal penal code is given by the policy $\hat{\pi} = \hat{\pi}^1 = \hat{\pi}^2$ in which the action profile is (a_2, b_2) is played in both states. For $\delta > 2/5$ this penal code is not an equilibrium, because the players prefer deviating from it in both states.

As an example of an induced equilibrium policy let us consider the non-stationary Markov policy π in which $\mu^k(x_2) = (a_1, b_1)$ for all $k \geq 0$, and $\mu^k(x_1) = (a_1, b_2)$ for odd k and $\mu^k(x_1) = (a_2, b_1)$ for even k. The first player's expected payoffs at stage k are

$$v_1^k(\pi)(x_1) = \begin{cases} 5/(1-\delta^2), & k \text{ even}, \\ 5\delta/(1-\delta^2), & k \text{ odd}, \end{cases}$$

and

$$v_1^k(\pi)(x_2) = \begin{cases} 5\delta^2/[(2-\delta)(1-\delta^2)], & k \text{ even} \\ 5\delta/[(2-\delta)(1-\delta^2)], & k \text{ odd}. \end{cases}$$

When deviations lead to π^1, the first player has no incentive to deviate from π. By symmetry the second player has no incentive to deviate either. Hence, for $\delta \geq 2/5$ the simple strategy $\sigma(\pi, p^*)$ is an equilibrium.

4. Characterization of the Equilibrium Payoff Correspondence

4.1. The Fixed-Point Operator

Let W be a correspondence from X to $\mathbb{R}^n$, i.e., $W(x)$ is a subset of $\mathbb{R}^n$ for all $x \in X$. Moreover, $F(W)$ denotes the functions obtained by selecting the elements from W, i.e., $v \in F(W)$ means that $v(x) \in W(x)$ for all $x \in X$. These functions will be used as continuation payoffs. The set $F_i(W)$ refers to player i's continuation payoff functions obtained from W. To shorten the notation let us denote

$$T_i(y, x, v_i) = \mathbb{E}_z \left[u_i(y, x, z) + \delta_i v_i(f(y, x, z)) \right],$$

where $v_i \in F_i(W)$, $i \in I$. Moreover, $T(y, x, v)$ will denote the vector with these components for $v = (v_1, \ldots, v_n) \in F(W)$. Note that, when we take y and x, the continuation payoff function v in T needs to be defined for all states that can be reached from the state x by taking the action profile

y. When considering one shot deviations, the punishment payoffs should be defined for all possible states that can be reached after such deviations. With a slight abuse of notation, the continuation payoff function $v \in F(W)$ in T denotes a function that is defined for all relevant states in question.

We say that an action profile is incentive compatible for given continuation payoff function and punishment payoffs followed by unilateral deviations, if there are no profitable one shot deviations. As usual, y_{-i} denotes the actions taken by the other players than player i.

Definition 5. An action profile $y \in Y(x)$ is incentive compatible at $x \in X$ for a continuation payoff function $v \in F(W)$ and punishment payoffs $\bar{v}_i \in F_i(W)$, $i \in I$, if

$$T_i(y, x, v_i) \geq \max_{y_i' \in Y_i(x)} T\left((y_i', y_{-i}), x, \bar{v}_i\right) \text{ for all } i \in I.$$

In the following we shall utilize the extremal punishment payoffs of W, i.e., $v_i^-(x; W) = \inf\{v_i(x); v(x) \in W(x)\}$. If unilateral deviations are followed by the extremal punishments payoffs, we get all action profiles and continuation payoffs that lead to incentive compatibility for some punishment payoffs.

Remark 1. If $y \in Y(x)$ is incentive compatible at $x \in X$ for $v \in F(W)$ and $\bar{v}_i \in F(W)$, $i \in I$, then y is incentive compatible at x for v and $v_i^-(\cdot; W) \in F_i(W)$, $i \in I$.

In the following the set $IC(y, x, W)$ denotes continuation payoffs $v \in F(W)$ for which $y \in Y(x)$ is incentive compatible at x when the punishment payoffs are given by $v_i^-(x; W)$ for all $x \in X$, $i \in I$. It follows from Remark 1 that $IC(y, x, W)$ contains all continuation payoffs functions for which y is incentive compatible at x for some punishment payoffs. At state x the correspondence W generates a set of payoffs that are obtained with incentive compatible action profiles and the corresponding continuation payoffs. These generated payoffs at x are

$$B(W)(x) = \{T(y, x, v) : y \in Y(x), v \in IC(y, x, W)\}.$$

Before going into the main result for V let us make some observations on the properties of operator B. First, it is monotone in the sense that if $W^1 \subseteq W^2$, i.e., $W^1(x) \subseteq W^2(x)$ for all $x \in X$, then $B(W^1) \subseteq B(W^2)$. Another property is that B preserves compactness, i.e., $B(W)(x)$ is compact for all $x \in X$, when W is compact valued. These properties will play an important role in showing that V can be found by using the fixed-point iteration.

Lemma 1. *The operator B has the following properties.*

1. *If* $W^1 \subseteq W^2$ *then* $B(W^1) \subseteq B(W^2)$.
2. *If* W *is compact valued, then* $B(W)$ *is compact valued.*

Proof. The first result is obvious. Let us show the compactness. Recall that stage game payoffs belong to compact sets, and $F(W)$ is compact. The latter follows from the compactness of $W(x)$ for all $x \in X$ and finiteness of X. If we pick a convergent sequence of payoffs in $B(W)(x)$, it then follows that there is a convergent subsequence $\{v^j(x)\}_j$ with the following properties. First,

$$v_i^j(x) = \mathbb{E}_z \left[u_i \left(y^j, x, z \right) + \delta_i \bar{v}_i^j \left(f \left(y^j, x, z \right) \right) \right], \ i \in I,$$

where $\bar{v}^j \in IC(y, x, W)$ for all $j \geq 0$. Second, $u(y^j, x, z)$, $x \in X$, $z \in Z$, and $\bar{v}^j$ converge as $j \to \infty$. Corresponding to the limit there is $y \in Y(x)$ such that $\lim_{j\to\infty} \bar{v}^j \in IC(y, x, W)$. Note that incentive compatibility holds in the limit. Thus, $B(W)(x)$, $x \in X$, are compact. □

4.2. Fixed-Point Equation for the Equilibrium Payoff Correspondence

The following result is analogous to self-generation for repeated games [4, 13]. Proposition 2 below says that if W generates itself at all states, i.e., $W(x) \subseteq B(W)(x)$ for all x, then the generated set $B(W)(x)$ is a subset of SPE payoffs. This result will be referred to as self-generation.

Proposition 2. *Let* W *be compact-valued correspondence such that* $W(x) \neq \emptyset$ *for all* $x \in X$. *Then* $W(x) \subseteq B(W)(x)$ *for all* x *implies* $B(W)(x) \subseteq V(x)$ *for all* x.

Proof. The proof proceeds as follows. We construct a policy corresponding to $v^0 \in F(W)$, then policies corresponding to extremal payoffs $v_i(\cdot; V)$, $i \in I$. This leads to a simple strategy. Finally we argue that the resulting simple strategy is an SPE.

Let us take $v^0 \in F(W)$. It follows from the assumption that for any $x^0 \in X$ we have $v^0(x^0) \in B(W)(x^0)$. Consequently, there are $y^0 \in Y(x^0)$ and $v^1 \in IC(y, x^0, W)$ such that $v^0(x^0) = T(y^0, x^0, v^1)$. Let us set $\mu^0(x^0) = y^0$ and repeat the argument for all initial points to get $\mu^0(x) \in Y(x)$ for all $x \in X$. Let us take any $x^0 \in X$ and $h^1 \in H^1(x^0)$, where $y^0 = \mu^0(x^0)$ is played, i.e., $h^1 = x^0 \cup \{y^0, z^0\}$. Because $v^1(x(h^1)) \in W(x(h^1))$, the assumption of the proposition implies that $v^1(x(h^1)) \in B(W)(x(h^1))$. We can now construct $\mu^1(x^0, z^0)$ similarly as we obtained μ^0. By repeating the argument at all stages we get a policy $\pi \in \Pi$ that the players follow as long as none of them deviates, i.e., an induced equilibrium policy.

Let us pick any $h^k \in H^k(\pi)$. Let $v^k(\pi)(x)$ denote the expected payoff when the players start to follow $\pi = (\mu^0, \mu^1, \ldots) \in \Pi$ from state $x \in X$ at

stage k. By definition we have

$$v_i^k(\pi)\left(x(h^k)\right) = T_i\left(\mu^k(h^k), x\left(h^k\right), v_i^{k+1}(\pi)\right) \text{ for all } i \in I,$$

where $h^{k+1} = h^k \cup \{\mu^k(h^k), z^k\} \in H^{k+1}(\pi)$. When the players follow π, i.e., there are no deviations at stage k, we get

$$v_i^k\left(x\left(h^k\right)\right) - v_i^k(\pi)\left(x(h^k)\right) = \delta_i \mathbb{E}\left[v_i^{k+1}\left(x\left(h^{k+1}\right)\right) - v_i^{k+1}(\pi)(x(h^{k+1}))|\pi\right]$$

for all $i \in I$. The expectation is over states conditional on the action profile given by π. By repeating the same deduction m times we get

$$v_i^k\left(x\left(h^k\right)\right) - v_i^k(\pi)\left(x(h^k)\right) = \delta_i^m \mathbb{E}\left[v_i^{k+m}\left(x\left(h^{k+m}\right)\right) - v_i^{k+m}(\pi)\left(x(h^{k+m})\right)|\pi\right]$$

for $i \in I$, when the players have followed π. The payoffs $v_i^{k+m}(x(h^{k+m}))$ and $v_i(\pi)(x(h^{k+m}))$, $i \in I$, are bounded because the stage-game payoffs u_i, $i \in I$, are bounded. It follows that

$$\delta_i^m\left[v_i^{k+m}\left(x\left(h^{k+m}\right)\right) - v_i(\pi)\left(x(h^{k+m})\right)\right] \to 0 \text{ when } m \to \infty.$$

Hence, $v^k(\pi)(x(h^k)) = v^k(x(h^k))$ for all $k \geq 0$ and $h^k \in H^k(\pi)$. In particular, it holds that $v_i(\pi)(x) = v_i^0(\pi)(x) = v_i^0(x)$ for all $x \in X$ and $i \in I$.

The above construction can be made especially for $v_i^-(\cdot; W) \in F_i(W)$, $i \in I$. Note that, $v_i^-(x; W)$, $i \in I$ are attained because W is compact valued. This follows from the arguments used in the proof of Proposition 1. Hence, we get policies $\pi^1, \ldots, \pi^n$ corresponding to the the extremal payoffs of W. Let p denote the resulting penal code. The penal code is an equilibrium because there are no profitable one shot deviations when the punishment payoffs are $v_i^-(\cdot; W)$, $i \in I$. By the incentive compatibility of π for the punishment payoffs given by p, there are no profitable one shot deviations from π either. Hence, by the one-shot deviation principle it holds that $\sigma(\pi, p) \in \Sigma$. □

The following proposition states the main result of the paper; SPE payoffs are characterized by a fixed-point of B. The result follows from the self-generation (Proposition 2), Proposition 1, and Lemma 1. This result is analogous to the Bellman equation [6] in dynamic programming; V takes the place of the value function and B corresponds to the Bellman operator.

Proposition 3. *V is the largest correspondence in set-inclusion that satisfies $V(x) = B(V)(x)$ for all x.*

Proof. By Proposition 2 it is enough to show that $V(x) \subseteq B(V)(x)$ for all x. Take an arbitrary $x \in X$ and $v^0(x) \in V(x)$. Then there is $\pi = (\mu^0, \mu^1, \ldots) \in \text{ind}(\Sigma)$ such that $v(\pi)(x^0) = v^0(x) = T(\mu^0(x), x, v^1(\pi))$. By Proposition 1 it

holds that $\sigma(\pi, p^*) \in \Sigma$, which means that $v^1(\pi) \in IC(\mu^0(x), x, V)$. Hence, $v^0(x) \in B(V)(x)$ and consequently $V(x) \subseteq B(V)(x)$. The monotonicity of B (Lemma 1) implies that V is the largest fixed-point of B in terms of set inclusion. □

Note that there may be several correspondences W satisfying $W = B(W)$. However, it follows from Proposition 2 that payoffs $W(x)$, $x \in X$, of such correspondences are obtained with SPE strategies, i.e., $W \subseteq V$.

5. Computation of the Equilibrium Payoff Correspondence

According to Proposition 3, to find V we need to find the largest fixed-point of B. In this section it is shown that the fixed-point iteration $W^{k+1} = B(W^k)$, i.e.,

$$W^{k+1}(x) = B(W^k)(x) \text{ for all } x \in X. \tag{2}$$

can be used for this purpose. Note that this iteration is analogous to the value iteration in dynamic programming. It will be shown that this iteration converges to the correspondence V, when the initial correspondence W^0 is sufficiently rich. For related works on computation of repeated game equilibria see [12] and [15]. These papers assume public randomization, which makes V and W^k, $k \geq 0$, convex-valued. Here public randomization is not assumed.

In the assumptions of the following results $m_i(x)$ is the min-max payoff for player i at state x, and $M_i(x)$ is maximum, respectively. Observe that the interval $[m_i(x), M_i(x)]$ contains all the payoffs that player i can get for initial state x. In the proof of convergence we shall need the following result.

Lemma 2. *When $\times_i[m_i(x), M_i(x)] \subseteq W^0(x)$ for all $x \in X$, then*

$$V(x) \subseteq W^k(x) \text{ and } W^{k+1}(x) \subseteq W^k(x) \text{ for all } x \in X,\ k \geq 0. \tag{3}$$

Proof. Let us first consider the case when $k = 0$. Take $v(x) \in B(W^0)(x)$. Then there is $\bar{v}^1 \in F(W^0)$ such that

$$v_i(x) = \mathbb{E}_z \left[u_i(y, x, z) + \delta_i \bar{v}_i^1(f(y, x, z)) \right],\ i \in I,$$

and $\bar{v}^1 \in IC(y, x, W^0)$. The condition $m_i(x) \leq \bar{v}_i^1(x) \leq M_i(x)$ for all $x \in X$ implies that $v_i(x) \in \times_i[m_i(x), M_i(x)]$, i.e., $v \in W^0(x)$. Hence, $W^1(x) = B(W^0)(x) \subseteq W^0(x)$.

Now we make the induction assumption $W^{k+1}(x) \subseteq W^k(x)$ for all $x \in X$. By the monotonicity of B we get

$$W^{k+2}(x) = B(W^{k+1})(x) \subseteq W^{k+1}(x) = B(W^k)(x) \text{ for all } x \in X. \tag{4}$$

Since $W^0(x)$ contains all the attainable payoffs when the initial state is x, $W^0(x)$ also contains $V(x)$, which implies $B(V)(x) \subseteq B(W^0)(x)$. Let us make the induction assumption that $V(x) \subseteq W^k(x)$ for all $x \in X$. By monotonicity of B it holds that $B(V)(x) \subseteq B(W^k)(x)$ for all $x \in X$. Due to self-generation of V we get the condition $V(x) \subseteq W^{k+1}(x)$ for all $x \in X$. □

Let us now show that W^k converges to V pointwise, i.e., $W^k(x) \to V(x)$ as $k \to \infty$ for all $x \in X$. The convergence of $W^k(x)$ to $V(x)$ refers to Painlevé-Kuratowski convergence, see, e.g., [5] and [21];

$$\liminf_{k\to\infty} W^k(x) = \limsup_{k\to\infty} W^k(x) = V(x).$$

Proposition 4. *Assume that W^0 is compact-valued and satisfies the assumptions of Lemma 2. Then $W^k(x) \to V(x)$ for all $x \in X$ as $k \to \infty$.*

Proof. The above lemmas imply the convergence, and

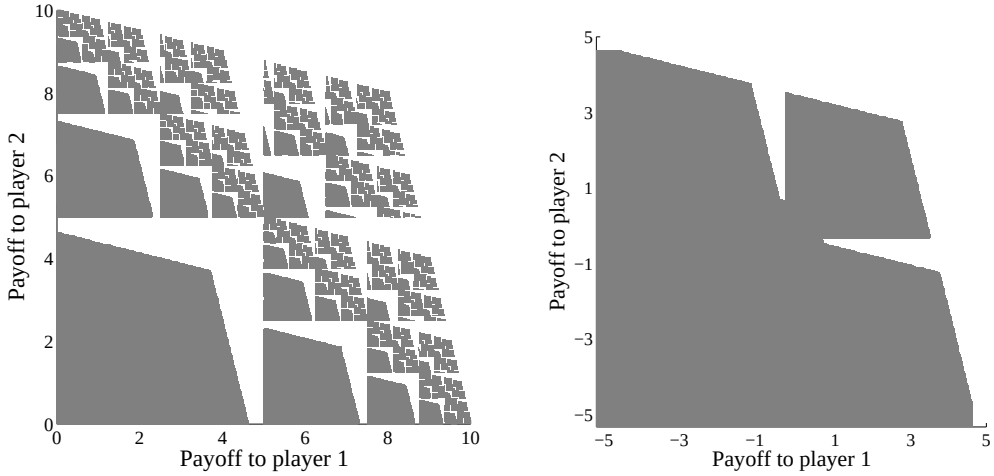
$$W(x) = \lim_{k\to\infty} W^k(x) = \cap_k W^k(x).$$

Let us assume that $W(x) \neq \emptyset$. Then by Lemma 2 we know that if $v(x) \in W(x)$, there are an action profile y and $\bar{v}$ such that $\bar{v} \in IC(y, x, W^k)$ for all $k \geq 0$, and $v_i(x) = \mathbb{E}[u_i(y, x, z) + \delta_i \bar{v}_i(f(y, x, z))]$, $i \in N$. This implies self-generation, i.e., $W(x) \subseteq B(W)(x)$ for all $x \in X$. By Proposition 2 we have $W(x) \subseteq V(x)$ for all $x \in X$. On the other hand, Lemma 2 implies that $V(x) \subseteq W(x)$ for all $x \in X$. Hence, we obtain $W = V$. □

Example 1. *This example continues the one presented in Section 3.3. The common discount factor δ is set to $1/2$. Approximations of sets $V(x_1)$ and $V(x_2)$ are presented in Figure 1. These sets are obtained by applying the iteration (2) with the initial correspondence $W^0(x) = \times_i [m_i(x), M_i(x)]$, $x \in X$.*

6. Finite Games and Behavior Strategies

In this section it will be assumed that $Y(x)$, $x \in X$, are finite sets. The purpose is to point out how the results of previous section can be extended to the case of behavior strategies. A behavior strategy maps histories of past actions and disturbances into distributions over actions profiles. The main benefit of such strategies is that the set of equilibria is non-empty.



a) Payoff set $V(x_1)$ b) Payoff set $V(x_2)$

Figure 1: Approximate equilibrium payoffs.

In the following $\Delta Y_i(x)$ denotes player i's mixed strategies at state $x \in X$. With slightly abusing the notation, $\Delta Y(x)$ is the collection of probability distributions $\alpha = (\alpha_1, \ldots, \alpha_n) \in \Delta Y(x)$ such that $\alpha_i \in \Delta Y_i(x)$. When assuming that players are using behavior strategies, we need perfect monitoring in the sense that the probability distributions used by the players are observed. Consequently, deviations from the ongoing policy can be punished. Formally, the set of histories should be appended with the players' past distributions over pure actions. As for pure strategies, we can argue that the players' equilibrium behavior is determined by policies.

The existence of subgame perfect equilibria in behavior strategies follows from earlier results for stochastic games. Beginning from the seminal work of [22] for zero-sum games, there is a plethora of existence results. Most notably, Mertens and Parthasarathy [19, 20] show the existence for games with general state and action sets. The technique used in their proof, and also in [18] and [24], is based on iterating a correspondence of equilibrium payoffs: W^{k+1} is the correspondence of Nash equilibrium payoffs in one-shot games with payoffs $T(y, x, v)$, where v belongs to $F(W^k)$. Clearly, this iteration is different from the iteration (2).

Assuming perfect monitoring and that $Y(x)$, $x \in X$, are finite, we can replicate the findings for pure strategies in the case of behavior strategies. In particular we obtain the following results.

Proposition 5. *The equilibrium correspondence V has the following proper-*

ties:

1. *$V(x) \neq \emptyset$ for all $x \in X$,*
2. *V is the largest fixed-point of operator B, and*
3. *the iteration (2) converges to V when W^0 is compact-valued and chosen as in Lemma 2.*

Proof. The proof follows by repeating the arguments used for pure strategies. Let us briefly sketch the differences for behavior strategies.

In the definition of B, the term $T(y, x, v)$ is replaced with $\mathbb{E}_\alpha[T(y, x, v)]$, where the expectation is over pure actions when $\alpha \in \Delta Y(x)$. Moreover, the incentive compatibility condition of player i in the definition of $B(W)(x)$ is then

$$\mathbb{E}_\alpha \left[T_i(y, x, v_i)\right] \geq \sup_{\alpha_i' \in \Delta Y_i(x)} \mathbb{E}_{\alpha_i', \alpha_{-i}} \left[T_i(y, x, v_i^-(\cdot; W))\right]. \tag{5}$$

In the proof of compactness of V, we can pick convergent subsequence such that the probabilities of pure actions at all stages converge. This can be done because $Y(x)$, $x \in X$, are finite. Moreover, it is worth noticing that the one-shot deviation principle holds for behavior strategies. Hence, we obtain the last two results in the same way as for pure strategies. Moreover, the results of Lemma 1 are valid.

As mentioned, the existence of equilibria, i.e., non-emptiness of follows from earlier results. However, let us briefly sketch how it is obtained by using the other two results. When W^0 is chosen as in Lemma 2, we can argue that then W^k will all become compact-valued and non-empty. The first follows from Lemma 1 and the latter by observing that when picking $v \in W^k$, $k \geq 0$, there are Nash equilibria in randomized strategies for the one-shot games with payoffs $T(y, x, v)$. These Nash equilibria satisfy the incentive compatibility condition (5), and hence $W^{k+1} = B(W^k)$, $k \geq 0$, are non-empty valued. The limit of the iteration (2), which is V, is then non-empty for all x. □

References

[1] D. Abreu, Extremal equilibria of oligopolistic supergames, Journal of Economic Theory 39 (1986) 191–225.

[2] D. Abreu, On the theory of infinitely repeated games with discounting, Econometrica 56 (1988) 383–396.

[3] D. Abreu, D. Pearce, E. Stacchetti, Optimal cartel equilibria with imperfect monitoring, Journal of Economic Theory 39 (1986) 251–269.

[4] D. Abreu, D. Pearce, E. Stacchetti, Toward a theory of discounted repeated games with imperfect monitoring, Econometrica 58 (1990) 1041–1063.

[5] J.P. Aubin, H. Frankowska, Set-Valued Analysis, Birkhäuser, Boston, 1990.

[6] R.E. Bellman, Dynamic Programming, Princeton University Press, Princeton N. J., 1957.

[7] K. Berg, M. Kitti, Equilibrium paths in discounted supergames, Working paper (2012).

[8] K. Berg, M. Kitti, Fractal geometry of equilibrium payoffs in discounted supergames, Working paper (2012).

[9] K. Berg, M. Kitti, Computing equilibria in discounted 2×2 supergames, Computational Economics 41 (2013) 71–88.

[10] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control: The Discrete Time Case, Athena Scientific, Belmont, Massachusetts, 1996.

[11] D. Blackwell, Discounted dynamic programming, Annals of Mathematical Statistics 36 (1965) 226–235.

[12] M.B. Cronshaw, Algorithms for finding repeated game equilibria, Computational Economics 10 (1997) 139–168.

[13] M.B. Cronshaw, D.G. Luenberger, Strongly symmetric subgame perfect equilibria in infinitely repeated games with perfect monitoring, Games and Economic Behavior 6 (1994) 220–237.

[14] P.K. Dutta, R. Radner, A game-theoretic approach to global warming, in: S. Kasuoka, A. Yamazaki (Eds.), Advances in Mathematical Economics, volume 8, Springer-Verlag, Tokyo, 2006, pp. 135–153.

[15] K. Judd, Ş. Yeltekin, J. Conklin, Computing supergame equilibria, Econometrica 71 (2003) 1239–1254.

[16] M. Kitti, Conditionally stationary equilibria in discounted dynamic games, Dynamic Games and Applications 1 (2011) 514–533.

[17] M. Kitti, Conditional Markov equilibria in discounted dynamic games, Mathematical Methods of Operations Research (to appear) (2013).

[18] A.P. Maitra, W.D. Sudderth, Subgame-perfect equilibria for stochastic games, Mathematics of Operations Research 32 (2007) 711–722.

[19] J.F. Mertens, T. Parthasarathy, Nonzerosum stochastic games, in: T.E.S. Raghavan, T.S. Ferguson, T. Parthasarathy, O.J. Vrieze (Eds.), Stochastic Games and Related Topics, Kluwer Academic Publishers, Boston, 1991.

[20] J.F. Mertens, T. Parthasarathy, Equilibria for discounted stochastic games, in: A. Neyman, S. Sorin (Eds.), Stochastic Games and Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003, pp. 131–172.

[21] R.T. Rockafellar, R.J.B. Wets, Variational Analysis, Springer, Berlin, 1998.

[22] L.S. Shapley, Stochastic games, Proceedings of the National Academy of Sciences of the USA 39 (1953) 1095–1100.

[23] C. Sleet, Ş. Yeltekin, On the computation of value correspondences, Working paper (2003).

[24] E. Solan, Discounted stochastic games, Mathematics of Operations Research 23 (1998) 1010–1021.

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133