A Direct Proof of the Existence of Correland Equilibrium Policies in General-Sum Markov Games

Amy Greenwald and Martin Zinkevich

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-05-07
July 2005

Brown University Technical Report (2005) CS-05-07

A Direct Proof of the Existence of Correlated Equilibrium Policies in General-Sum Markov Games

Amy Greenwald AMY@BROWN.EDU
Department of Computer Science
Brown University
Providence, RI 02912

Martin Zinkevich MAZ@CS.BROWN.EDU
Department of Computer Science
Brown University
Providence, RI 02912

Abstract

In this report, we establish the existence of stationary correlated equilibrium policies in Markov games. In fact, the existence of stationary correlated equilibrium policies in Markov games is implied by the existence of stationary Nash equilibrium policies in Markov games (see, for example, Filar and Vrieze (1996)). Nonetheless, we prove existence directly because our existence proof, a fixed point argument, motivates our algorithmic search for just such a fixed point in a companion paper (Greenwald et al., 2005). Analogously, a classic proof of existence of an optimal policy in Markov decision processes, based on Banach's fixed point theorem, motivates the design of value iteration (see, for example, Puterman (1994)) and Q-learning Watkins (1989).

Keywords: Multiagent Learning, Reinforcement Learning, Markov Games

1. Correlated Equilibrium Policies in Markov Games

We begin with some notation and terminology that we rely on to define Markov games. We adopt the following standard game-theoretic terminology: the term action (strategy, or policy) *profile* is used to mean a vector of actions (strategies, or policies), one per player. In addition, $\Delta(X)$ denotes the set of all probability distributions over finite set X.

Definition 1 *A (finite) discounted **Markov game** is a tuple $\Gamma_\gamma = \langle N, S, A, P, R\rangle$ in which*

- *N is a finite set of n players*
- *S is a finite set of m states*
- *$A = \prod_{i\in N, s\in S} A_i(s)$, where $A_i(s)$ is player i's finite set of pure actions at state s; we define $A(s) \equiv \prod_{i\in N} A_i(s)$ and $A_{-i}(s) = \prod_{j\neq i} A_j(s)$, so that $A(s) = A_{-i}(s) \times A_i(s)$; we write $a = (a_{-i}, a_i) \in A(s)$ to distinguish player i, with $a_i \in A_i(s)$ and $a_{-i} \in A_{-i}(s)$; we also define $\mathcal{A} = \bigcup_{s\in S} \bigcup_{a\in A(s)} \{(s, a)\}$, the set of state-action pairs.*

- P *is a system of transition probabilities: i.e., for all* $s \in S$, $a \in A(s)$, $P[s' \mid s, a] \geq 0$ *and* $\sum_{s' \in S} P[s' \mid s, a] = 1$*; we interpret* $P[s' \mid s, a]$ *as the probability that the next state is* s' *given that the current state is* s *and the current action profile is* a
- $R : \mathcal{A} \to [\alpha, \beta]^n$*, where* $R_i(s, a) \in [\alpha, \beta]$ *is player* i*'s reward at state* s *and at action profile* $a \in A(s)$
- $\gamma \in [0, 1)$ *is a discount factor*

Let us imagine that in addition to the players, there is also a *referee*,[1] who can be considered to be a physical machine (i.e., the referee itself has no beliefs, desires, or intentions). At each time step, the referee sends to each player a private signal consisting of a recommended action for that player.[2] We assume the referee selects these actions according to a *stationary* policy $\pi \in \prod_{s \in S} \Delta(A(s))$: i.e., a policy that depends only on state, not on time.

The dynamics of a discrete-time Markov game *with a referee* unfold as follows: at time $t = 1, 2, \ldots$, the players and the referee observe the current game state $s^t \in S$; following its policy π, the referee selects the distribution π_{s^t}, based on which it recommends an action, say α_i^t, to each player i; given its recommendation, each player selects an action a_i^t, and the pure action profile $a^t = (a_1^t, \ldots, a_n^t)$ is played; based on the current state and action profile, each player i now earns reward $R_i(s^t, a^t)$; finally, nature selects a successor state s^{t+1} with transition probability $P[s^{t+1} \mid s^t, a^t]$; the process repeats at time $t + 1$.

Note that at each stage of the game, players are aware of the history of all past states and all past actions. We assume the structure of the Markov game, including rewards and transitions, is common knowledge to the players, so that they can infer past rewards and expected future transitions from information about past states and actions.

In analyzing the players' motivations, we assume that their behavior is rational. Following von Neumann and Morgenstern (1944) utility theory, we assume that players use all of the information at their disposal—the history of states and actions, as well as the history of the signals received from the referee—to maximize their expected utility.

Do rational players in a Markov game follow the advice of the referee? Not surprisingly, the answer to this question depends upon the referee's policy π. We derive necessary and sufficient conditions for π to be a correlated equilibrium: that is, for π to consist of recommendations that rational players are motivated to follow.

Throughout the rest of this section, we rely on basic fixed point theory. A review of the standard terminology and fundamental theorems is included in Appendix A.

1. Note that the referee is not part of the definition of a Markov game. While a referee can be of assistance in the implementation of a correlated equilibrium, the concept can be defined without reference to this third party. In this section, we introduce the referee as a pedagogical device. In our experimental work (Greenwald et al., 2005), we sometimes rely on the referee to facilitate the implementation of correlated equilibria.
2. Generalizing sunspot equilibria (Shell, 1989), which rely on public randomization devices, to define or implement a correlated equilibria, the referee sends a *private*, rather than a public, signal to each player. It suffices for the referee to send to each player as this private signal precisely its recommended action. Any joint distribution of the players' actions that could arise by the referee sending more general signals can also be achieved by the referee sending each player its recommended action (Aumann, 1974).

1.1 Correlated Equilibrium in One-Shot Games: A Review

A (finite) *one-shot game* is a tuple $\Gamma = \langle N, A, R \rangle$ in which N is a finite set of n players; $A = \prod_{i \in N} A_i$, where A_i is player i's finite set of pure actions; and $R : A \to \mathbb{R}^n$, where $R_i(a)$ is player i's reward at action profile $a \in A$.

Once again, imagine a referee who selects an action profile a according to some policy $\pi \in \Delta(A)$. The referee advises player i to follow action a_i. Define $A_{-i} = \prod_{j \neq i} A_j$. Define $\pi(a_i) = \sum_{a_{-i} \in A_{-i}} \pi(a_{-i}, a_i)$ and $\pi(a_{-i} \mid a_i) = \frac{\pi(a_{-i}, a_i)}{\pi(a_i)}$ whenever $\pi(a_i) > 0$.

Definition 2 *Given a one-shot game Γ, the policy $\pi \in \Delta(A)$ is a **correlated equilibrium** if, for all $i \in N$, for all $a_i \in A_i$ with $\pi(a_i) > 0$, and for all $a'_i \in A_i$,*

$$\sum_{a_{-i} \in A_{-i}} \pi(a_{-i} \mid a_i) R_i(a_{-i}, a_i) \geq \sum_{a_{-i} \in A_{-i}} \pi(a_{-i} \mid a_i) R_i(a_{-i}, a'_i) \tag{1}$$

If the referee chooses a according to a correlated equilibrium, then the players are motivated to follow his advice, because the expression $\sum_{a_{-i} \in A_{-i}} \pi(a_i \mid a_i) R(a_{-i}, a'_i)$ computes player i's expected reward for playing a'_i when the referee advises him to play a_i.

Equivalently, for all $i \in N$ and for all $a_i, a'_i \in A_i$,

$$\sum_{a_{-i} \in A_{-i}} \pi(a_{-i}, a_i) R_i(a_{-i}, a_i) \geq \sum_{a_{-i} \in A_{-i}} \pi(a_{-i}, a_i) R_i(a_{-i}, a'_i) \tag{2}$$

Equation 2 is Equation 1 multiplied by $\pi(a_i)$. Equation 2 holds trivially whenever $\pi(a_i) = 0$, because in such cases both sides equal zero. Given a one-shot game Γ, $R(a_{-i}, a_i)$ is known, which implies that Equation 2 is a system of linear inequalities, with $\pi(a_{-i}, a_i)$ unknown.

The set of all solutions to a system of linear inequalities is convex. Since these inequalities are not strict, this set is also closed. This set is bounded as well, because the set of all policies is bounded. Therefore, the set of correlated equilibria is compact and convex.

If the recommendations of the referee in a correlated equilibrium are independent (i.e., for all $i \in N$, for all $a_i, a'_i \in A_i$, for all $a_{-i} \in A_{-i}$, $\pi(a_{-i} \mid a_i) = \pi(a_{-i} \mid a'_i)$, whenever $\pi(a_i), \pi(a'_i) > 0$), then a correlated equilibrium is also a Nash equilibrium. In fact, any Nash equilibrium can be represented as a correlated equilibrium: the players can simply generate their own advice (independently). Existence of both types of equilibria is ensured by Nash's theorem (Nash, 1951). Therefore, the set of correlated equilibria is nonempty, as well. We have established the following (well-known) result.

Theorem 3 *The set of correlated equilibria in a one-shot game is nonempty, compact, and convex.*

Finally, we note that a correlated equilibrium in a one-shot game can be computed in polynomial time via linear programming. Equation 2 consists of $\sum_{i \in N} |A_i| (|A_i| - 1)$ linear inequalities, which is polynomial in the number of players, and $\prod_{i \in N} |A_i| - 1$ variables, which is exponential in the number of players, but polynomial in the size of the game.

Game 1: Bach or Stravinsky

	B	S
b	2,1	0,0
s	0,0	1,2

1.1.1 Examples

Here, we show by example some of the benefits of correlated equilibria over Nash equilibria. Game 1 represents a situation where two agents, the first a Bach lover and the second a Stravinsky lover, are deciding whether to go to a Bach concert or a Stravinsky concert. The first agent selects a row, and the second agent selects a column. The utility profile of each outcome is written in each cell. The first agent prefers to go to the Bach concert, but given a choice between going with the second agent to the Stravinsky concert or going to the Bach concert alone, she prefers the former. Similarly, the second agent prefers to go to the Stravinsky concert, but given a choice between going with the first agent to the Bach concert or going to the Stravinsky concert alone, he too prefers the former.

In Game 1, there are two pure strategy Nash equilibria: one where the agents play (b, B) (go to a Bach concert together) and another where the agents play (s, S) (go to a Stravinsky concert together). Either of these behaviors could be considered "unfair:" the outcome (s, S) is unfair to the first agent, since she would rather go to the Bach concert; similarly, the outcome (b, B) is unfair to the second agent, since he would rather go to the Stravinsky concert. There is one mixed strategy Nash equilibrium in this game, where the first agent plays b with probability $2/3$ and the second agent plays s with probability $2/3$. Here, each agent obtains a utility of $2/3$, which is "fair," but is also less than either agent would obtain were either of the pure strategy Nash equilibria to be played.

If two people were seeking a fair solution to this game, they might decide to flip a (fair) coin, agreeing in advance that if the coin comes up heads, they both go to the Bach concert, whereas if the coin comes up tails, they both go to the Stravinsky concert. This solution is an example of a correlated equilibrium, where $\pi(b, B) = 1/2$ and $\pi(s, S) = 1/2$. Not only is this solution fair, it is also Pareto-optimal, that is, no agent can be made better off without making some other agent worse off.[3]

3. In Markov games, such a balance can be achieved another way. For instance, in a Markov game with Game 1 as the only state (i.e., the infinitely repeated game of Bach or Stravinsky), the agents can alternate between going to the Bach concert and going to the Stravinsky concert. However, it is not always the case that time can be used as a correlation device, as the following example shows.

   Imagine three agents choosing a number between 1 and 100. If the first two agents choose the same number, but the third agent chooses a different number, then the third agent must pay each of the first two agents one dollar. Otherwise, each of the first two agents pay the third agent one dollar. A game in this spirit is played by a quarterback (the first agent), a wide receiver (the second agent), and an opposing cornerback (the third agent): the quarterback and the wide receiver must agree upon a passing pattern which the opposing cornerback must guess.

   Now, observe that it is a Nash equilibrium for each agent to choose a number uniformly at random between 1 and some $m \in \{1, \ldots, 100\}$. Among these equilibria, the one that is best for the first two agents is $m = 2$, in which case they win a quarter of the time. However, if they correlate their actions, both of them choosing the same number k uniformly at random between 1 to 100, they win 99/100 of the time. In this game, the agents cannot use time as a correlation device because the third agent could anticipate their choices.

Game 2: Shapley's Game

	R	P	S
r	0,0	0,1	1,0
p	1,0	0,0	0,1
s	0,1	1,0	0,0

In Shapley's Game (Game 2), both agents earn a higher utility by playing a correlated equilibrium instead of a Nash equilibrium. (Shapley's game differs from Rock-Paper-Scissors only in that in the latter the diagonal entries are $(\frac{1}{2}, \frac{1}{2})$.) At the unique Nash equilibrium, each agent chooses an action uniformly at random and each agent's expected utility is 1/3. However, if a referee selects an action profile uniformly at random from the set

$$\{(r, P), (r, S), (p, R), (p, S), (s, R), (s, P)\}$$

and if the two agents follow the referee's advice, then each agent's expected utility is 1/2. Initially, one might think that the referee could select uniformly at random from, say $\{(r, P), (p, R)\}$. But then, if the first agent were advised to play r, she could infer that the second agent was advised to play P, which would motivate her to play s. If one agent were to cooperate with the referee, the other agent would be motivated to deviate.

By playing a correlated equilibrium in the Bach or Stravinsky game, the agents achieve a fair and Pareto-optimal solution. In Shapley's game, by playing a correlated rather than a Nash equilibrium, all of the agents fare better. In addition, a correlated equilibrium in a one-shot game can be computed in polynomial time. The corresponding complexity question for Nash equilibrium is open, although it is known that it is NP-hard to compute certain classes of Nash equilibria (Gilboa and Zemel, 1989; Conitzer and Sandholm, 2003).

1.2 Correlated Equilibrium Policies in Markov Games: Definition

We are now ready to address the question of whether or not agents are willing to follow the advice of a referee in a Markov game. To do so, we compute the expected utility of an agent when it follows the advice of the referee as well as the expected utility of an agent when it deviates, in both cases assuming all other agents follow the advice of the referee.

Given a Markov game Γ_γ, and a referee's policy π, consider the transition matrix T^π such that $T^\pi_{ss'}$ is the probability of transitioning to state s' from state s, given that the referee selects an action profile according to the distribution π_s that the agents indeed follow:

$$T^\pi_{ss'} = \sum_{a \in A(s)} \pi_s(a) P[s' \mid s, a] \tag{3}$$

Exponentiating this matrix, the probability of transitioning to state s' from state s after t time steps is given by $(T^\pi_{ss'})^t$. Now the value function $V^\pi_i(s)$ represents agent i's expected reward, originating at state s, assuming all agents follow the referee's policy π:

$$V^\pi_i(s) = (1-\gamma) \sum_{t=0}^{\infty} \sum_{s' \in S} \gamma^t (T^\pi_{ss'})^t \sum_{a \in A(s')} \pi_{s'}(a) R_i(s', a) \tag{4}$$

The Q-value function $Q_i^\pi(s,a)$ represents agent i's expected rewards if action profile a is played in state s and the referee's policy π is followed thereafter:

$$Q_i^\pi(s,a) = (1-\gamma)\left(R_i(s,a) + \gamma \sum_{s' \in S} P[s'|s,a]\left(\sum_{t=0}^{\infty} \sum_{s'' \in S} \gamma^t (T_{s's''}^\pi)^t \sum_{a \in A(s'')} \pi_{s''}(a) R_i(s'',a)\right)\right) \tag{5}$$

The normalization constant $1-\gamma$ ensures that the ranges of V_i^π and Q_i^π each fall in $[\alpha,\beta]$.

The following theorem, which we state without proof, is an analog of Bellman's Theorem (Bellman, 1957) that follows directly from Equations 4 and 5 via the Markov property. (Note also that the referee's policy π is stationary by assumption.)

Theorem 4 *Given a Markov game Γ_γ, for any $V : S \to [\alpha,\beta]^n$, for any $Q : \mathcal{A} \to [\alpha,\beta]^n$, and for any stationary policy π, $V = V^\pi$ and $Q = Q^\pi$ if and only if:*

$$V_i(s) = \sum_{a \in A(s)} \pi_s(a) Q_i(s,a) \tag{6}$$

$$Q_i(s,a) = (1-\gamma) R_i(s,a) + \gamma \sum_{s' \in S} P[s'|s,a] V_i(s') \tag{7}$$

Hereafter, in place of Equations 4 and 5, we define V_i^π and Q_i^π recursively as the unique pair of functions satisfying Equations 6 and 7.

Definition 5 *Given a Markov game Γ_γ, a referee's policy π is a **correlated equilibrium** if for any agent i, if all the other agents follow the advice of the referee, agent i maximizes its expected utility by also following the advice of the referee.*

In this section, by assuming that the referee's policy is stationary, we restrict our attention to stationary correlated equilibrium policies.

Define $\pi_s(a_i) = \sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i}, a_i)$ and $\pi_s(a_{-i} \mid a_i) = \frac{\pi_s(a_{-i},a_i)}{\pi_s(a_i)}$ whenever $\pi_s(a_i) > 0$.

Remark 6 *Given a Markov game Γ_γ, a stationary policy π is* **not** *a correlated equilibrium if there exists an $i \in N$, an $s \in S$, an $a_i \in A_i(s)$ with $\pi(a_i) > 0$, and an $a_i' \in A_i(s)$,* s.t.*:*

$$\sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i} \mid a_i) Q_i^\pi(s,(a_{-i},a_i)) < \sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i} \mid a_i) Q_i^\pi(s,(a_{-i},a_i')) \tag{8}$$

Here, in state s, when it is recommended that agent i play a_i, it would rather play a_i', because the expected utility of a_i' is greater than the expected utility of a_i. This is an example of a *one-shot deviation* (see, for example, Osborne and Rubinstein (1994)). The definition of correlated equilibrium in Markov games, however, permits arbitrarily complex deviations on the part of an agent: e.g., deviations could be nonstationary. The next theorem states that the converse of Remark 6 is also true, implying that it suffices to consider one-shot deviations. Together Remark 6 and Theorem 7 provide the necessary and sufficient conditions for π to be a stationary correlated equilibrium policy in a Markov game.

Theorem 7 *Given a Markov game Γ_γ, a stationary policy π is a correlated equilibrium if for all $i \in N$, for all $s \in S$, for all $a_i \in A_i(s)$ with $\pi(a_i) > 0$, for all $a_i' \in A_i(s)$,*

$$\sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i} \mid a_i) Q_i^\pi(s, (a_{-i}, a_i)) \geq \sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i} \mid a_i) Q_i^\pi(s, (a_{-i}, a_i')) \tag{9}$$

Here, in state s, when it is recommended that agent i play a_i, it is happy to play a_i, because the expected utility of a_i is greater than or equal to the expected utility of a_i', for all a_i'.

Observe the following: if all of the other agents but agent i play according to the referee's policy π, then from the point of view of agent i, its environment is an MDP. Hence, the one-shot deviation principle for MDPs establishes Theorem 7 (see Appendix C).

Corollary 8 *Given a Markov game Γ_γ, a stationary policy π is a correlated equilibrium if for all $i \in N$, for all $s \in S$, and for all $a_i, a_i' \in A_i(s)$,*

$$\sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i}, a_i) Q_i^\pi(s, (a_{-i}, a_i)) \geq \sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i}, a_i) Q_i^\pi(s, (a_{-i}, a_i')) \tag{10}$$

Equation 10 is Equation 9 multiplied by $\pi_s(a_i)$.

Unlike in one-shot games where only the $\pi(a_{-i}, a_i)$'s are unknown (see Equation 2), here the $\pi_s(a_{-i}, a_i)$'s, and hence the $Q_i^\pi(s, (a_{-i}, a_i))$'s, are unknown. In particular, Equation 10 is not a system of linear inequalities, but rather a system of nonlinear inequalities.

1.3 Correlated Equilibrium Policies in Markov Games: Existence

In Theorem 4, Equation 6 is dependent upon the referee's policy π, whereas Equation 7 also depends on the structure of the game. Like value iteration and Q-learning, which are used to compute optimal policies in MDPs, we can try to leverage this separation to search for correlated equilibrium policies in Markov games.

Given a Markov game Γ_γ, with a particular S and $\mathcal{A}$, define the following three spaces:

1. $\mathcal{V} = [\alpha, \beta]^{n \times S}$, the space of all functions of the form $V : S \to [\alpha, \beta]^n$
2. $\mathcal{Q} = [\alpha, \beta]^{n \times \mathcal{A}}$, the space of all functions of the form $Q : \mathcal{A} \to [\alpha, \beta]^n$
3. $\Pi = \prod_{s \in S} \Delta(A(s))$, the space of all policies

When viewed as subsets of Euclidean space, these are nonempty, compact, convex sets.

Let us express the value function defined in Equation 6 more precisely by making explicit its dependence on the Q-values as well as the policy π. Define $V_{\mathcal{Q} \times \Pi} : \mathcal{Q} \times \Pi \to \mathcal{V}$ such that for all $Q \in \mathcal{Q}$, for all $\pi \in \Pi$, for all $i \in N$, and for all $s \in S$,

$$(V_{\mathcal{Q} \times \Pi}(Q, \pi))_i(s) = \sum_{a \in A(s)} \pi_s(a) Q_i(s, a) \tag{11}$$

Similarly, let us also express the Q-value function defined in Equation 7 more precisely so that its dependence on the value function is made explicit. Define $Q_\mathcal{V} : \mathcal{V} \to \mathcal{Q}$ such that for all $V \in \mathcal{V}$, for all $i \in N$, for all $s \in S$, and for all $a \in A(s)$,

$$(Q_\mathcal{V}(V))_i(s, a) = (1 - \gamma) R_i(s, a) + \gamma \sum_{s' \in S} P[s' | s, a] V_i(s') \tag{12}$$

Here, we have highlighted the fact that the dependence of Q on π arises through V. Finally, we define the correspondence $\pi^*_{\mathcal{Q}} : \mathcal{Q} \rightarrow\!\!\!\!\rightarrow \Pi$ such that for all $Q \in \mathcal{Q}$, $\pi \in \Pi$ is in $\pi^*_{\mathcal{Q}}(Q)$ if and only if for all $s \in S$, for all $a_i, a'_i \in A_i(s)$:

$$\sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i}, a_i) Q(s, (a_{-i}, a_i)) \geq \sum_{a_{-i} \in A_{-i}(s)} \pi_s(a_{-i}, a_i) Q(s, (a_{-i}, a'_i)) \tag{13}$$

Theorem 9 *Given a Markov game Γ_γ, for $V \in \mathcal{V}$, $Q \in \mathcal{Q}$, and $\pi \in \Pi$, if $V = V_{\mathcal{Q}\times\Pi}(Q, \pi)$, $Q = Q_{\mathcal{V}}(V)$, and $\pi \in \pi^*_{\mathcal{Q}}(Q)$, then $V = V^\pi$, $Q = Q^\pi$, and π is a stationary correlated equilibrium policy.*

Proof By Equations 11 and 12, for all $i \in N$, for all $s \in S$, for all $a \in A(s)$:

$$V_i(s) = \sum_{a \in A(s)} \pi_s(a) Q_i(s, a) \tag{14}$$

$$Q_i(s, a) = (1 - \gamma) R_i(s, a) + \gamma \sum_{s' \in S} P[s'|s, a] V_i(s') \tag{15}$$

Therefore, by Theorem 4, $V = V^\pi$ and $Q = Q^\pi$. Finally, since $\pi \in \pi^*_{\mathcal{Q}}(Q^\pi)$, it follows from Corollary 8 that π is a stationary correlated equilibrium policy. ■

Define $\rho^* \equiv (I \times \pi^*_{\mathcal{Q}}) \circ Q_{\mathcal{V}} \circ V_{\mathcal{Q}\times\Pi}$, where I is the identity function.

Corollary 10 *Given a Markov game Γ_γ, a fixed point (if it exists) of the correspondence $\rho^* : \mathcal{Q} \times \Pi \rightarrow\!\!\!\!\rightarrow \mathcal{Q} \times \Pi$ is a pair consisting of a stationary correlated equilibrium policy and its associated Q-values.*

Proof Choose (Q, π) to be a fixed point of ρ^*. Define $V = V_{\mathcal{Q}\times\Pi}(Q, \pi)$.

$$(Q, \pi) \in \rho^*(Q, \pi) \tag{16}$$

$$= \{(Q_{\mathcal{V}}(V_{\mathcal{Q}\times\Pi}(Q, \pi)), \pi') \mid \pi' \in \pi^*_{\mathcal{Q}}(Q_{\mathcal{V}}(V_{\mathcal{Q}\times\Pi}(Q, \pi)))\} \tag{17}$$

$$= \{(Q_{\mathcal{V}}(V), \pi') \mid \pi' \in \pi^*_{\mathcal{Q}}(Q_{\mathcal{V}}(V))\} \tag{18}$$

Therefore, $Q = Q_{\mathcal{V}}(V)$ so that $\pi \in \pi^*_{\mathcal{Q}}(Q)$. By Theorem 9, $V = V^\pi$, $Q = Q^\pi$, and π is a stationary correlated equilibrium policy. ■

Corollary 11 *A fixed point (if it exists) of the correspondence $Q^*_{\mathcal{Q}} : \mathcal{Q} \rightarrow\!\!\!\!\rightarrow \mathcal{Q}$ where*

$$Q^*_{\mathcal{Q}} \equiv \mathcal{Q}_{\mathcal{V}} \circ \mathcal{V}_{\mathcal{Q}\times\Pi} \circ (I \times \pi^*_{\mathcal{Q}}) \tag{19}$$

are the Q-values associated with a stationary correlated equilibrium policy.

The proof is analogous to the proof of Corollary 10.

Corollary 12 *A fixed point (if it exists) of the correspondence $V^*_{\mathcal{V}} : \mathcal{V} \to\to \mathcal{V}$ where*

$$V^*_{\mathcal{V}} \equiv V_{\mathcal{Q}\times\Pi} \circ (I \times \pi^*_{\mathcal{Q}}) \circ Q_{\mathcal{V}} \tag{20}$$

is the value function associated with a stationary correlated equilibrium policy.

The proof is analogous to the proof of Corollary 10.

Note that $Q^*_{\mathcal{Q}}$ has a fixed point if and only if $V^*_{\mathcal{V}}$ has a fixed point, and $V^*_{\mathcal{V}}$ has a fixed point if and only if ρ^* has a fixed point. In what follows, we prove that ρ^* has a fixed point.

Theorem 13 *If the correspondence $\pi_{\mathcal{Q}} : \mathcal{Q} \to\to \Pi$ is closed with nonempty, compact, convex values, then the correspondence $\rho = (I \times \pi_{\mathcal{Q}}) \circ Q_{\mathcal{V}} \circ V_{\mathcal{Q}\times\Pi}$ has a fixed point.*

We prove this theorem in Appendix B. The proof relies on Kakutani's Fixed Point Theorem, a generalization of Brouwer's Fixed Point Theorem.

To show that the correspondence ρ^* has a fixed point, we proceed by showing that the correspondence $\pi^*_{\mathcal{Q}}$ satisfies the requisite properties in Theorem 13.

Lemma 14 *The correspondence $\pi^*_{\mathcal{Q}}$ has nonempty, compact, convex values.*

Proof Fix $Q \in \mathcal{Q}$ and $s \in S$. The set $(\pi^*_{\mathcal{Q}}(Q))_s$ is the set of correlated equilibria of the one-shot game $\Gamma = \langle N', A', R' \rangle$ where $N' = N$, $A'_i = A_i(s)$, and for all $i \in N'$, $a \in A'$, $R'_i(a) = Q_i(s, a)$. By Theorem 3, $(\pi^*_{\mathcal{Q}}(Q))_s$ is a nonempty, convex, compact set. Therefore, $\pi^*_{\mathcal{Q}}(Q)$, which is the product of multiple nonempty, convex, compact sets, is itself a nonempty, convex, compact set. ■

To show that the correspondence $\pi^*_{\mathcal{Q}}$ is closed, we rely on the following technical lemma.

Lemma 15 *Given two polynomials $f : \mathbb{R}^r \to \mathbb{R}$, $g : \mathbb{R}^s \to \mathbb{R}$, the set of all points satisfying $f(x) \geq g(x)$ is closed.*

Proof Choose a sequence x^m such that $f(x^m) \geq g(x^m)$ for all m. For the sake of contradiction, assume that the sequence x^m converges to a point x^* where $f(x^*) < g(x^*)$; in particular, there exists an $\epsilon > 0$ such that $g(x^*) - f(x^*) > \epsilon$. Since $g - f$ is continuous, there exists a $\delta > 0$ such that for all x, if $d(x, x^*) < \delta$, then $d\left((g(x^*) - f(x^*)), (g(x) - f(x))\right) < \epsilon$. Because $x^m \to x^*$, there exists an N such that for all $n > N$, $d(x^n, x^*) < \delta$. But then, for all such n, $g(x^n) - f(x^n) > 0$, which implies that x^n does not satisfy $f(x^n) \geq g(x^n)$. Contradiction. ■

Lemma 16 *The correspondence $\pi^*_{\mathcal{Q}}$ is closed.*

Proof Observe that the graph of $\pi^*_{\mathcal{Q}}$ is the set of points satisfying the finite set of inequalities in Equation 13. Each of these inequalities is an inequality between two polynomials; thus, by Lemma 15, the set of points satisfying each inequality is closed. The set of points satisfying all of these inequalities is the intersection of a finite number of closed sets, which is itself closed. Since the graph of $\pi^*_{\mathcal{Q}}$ is a closed set, the correspondence $\pi^*_{\mathcal{Q}}$ is closed. ■

Theorem 17 *The correspondence* $\rho^* = (I \times \pi^*_{\mathcal{Q}}) \circ Q_{\mathcal{V}} \circ V_{\mathcal{Q}\times\Pi}$ *has a fixed point.*

Proof This result follows directly from Theorem 13, with $\pi_{\mathcal{Q}} = \pi^*_{\mathcal{Q}}$, which has the desired properties by Lemmas 14 and 16. ■

Theorem 18 *Every Markov game has a stationary correlated equilibrium policy.*

Proof By Theorem 17, the correspondence $\rho^* = (I \times \pi^*_{\mathcal{Q}}) \circ Q_{\mathcal{V}} \circ V_{\mathcal{Q}\times\Pi}$ has a fixed point. By Corollary 10, this fixed point is comprised of a stationary correlated equilibrium policy. ■

2. Related Work

There is a large amount of literature on Markov games, also known as stochastic games, dating back to Shapley's seminal paper (1953a), in which he demonstrated the existence of minimax equilibrium policies as well as the convergence of "minimax value iteration" in finite, discounted, two-player, zero-sum Markov games. The theory was extended to general-sum Markov games by Fink (1964), who demonstrated the existence of Nash equilibrium policies in such games. Like Nash's original proof of the existence of Nash equilibria in one-shot games (Nash, 1951), standard proofs of this result (see, for example, (Neyman and Sorin, 2003)) construct a best-response correspondence from policies to policies, the fixed points of which are Nash equilibrium policies. In contrast, we define the correspondence ρ^* from pairs of Q-functions and policies to pairs of Q-functions and policies, which is suggestive of an iterative algorithm for computing a fixed point, in the spirit of value iteration. The same can be said of the correspondences $Q^*_{\mathcal{Q}}$ from Q-functions to Q-functions and $V^*_{\mathcal{V}}$ from value functions to value functions.

Correlated equilibrium was first defined by Aumann (1974) for one-shot games. Under one interpretation, Aumann's definition relies on a **correlation device** (i.e., the referee), which sends private signals (i.e., recommendations) to the players before the game is played. This device was extended by Forges (1986) to a **communication device**, applicable to extensive form games (see, for example, Osborne and Rubinstein (1994)). A communication device allows for signals to be communicated to and from the device throughout the game. Solan and Vieille (2002) further extended this device, defining a **general communication device** that is aware of the history of play as well as past signals sent and received. Using this device, they generalize the notion of correlated equilibrium in Markov games. One special case of a general communication device is a **stationary** communication device (Solan and Vieille, 2002), which draws its signals from the same distribution at all times. In our definition of stationary correlated equilibrium policies in Markov games, we rely on stationary communication devices.[4]

4. Although we define the referee's policy as a function from states to distributions over actions, such stationary policies can be implemented as stationary communication devices: simply define signals as functions from states to distributions over actions, so that the referee can draw its signals from the same distribution at all times without knowledge of the current state.

3. Summary

In this report, we relied on Kakutani's fixed point theorem to establish the existence of stationary correlated equilibrium policies in Markov games. Specifically, we defined a correspondence, the fixed points of which are the stationary correlated equilibrium policies of a Markov game, and we showed that this correspondence satisfies Kakutani's sufficient conditions, ensuring that the set of such fixed points is nonempty.

The definition of this correspondence suggests an algorithm for computing one of its fixed points: given initial Q-values and an initial policy, update the values, Q-values, and policy, and repeat. In MDPs, the special case of Markov games with only a single agent, this procedure—known as value iteration—converges to a unique fixed point V^*, a unique fixed point Q^*, and a corresponding stationary, optimal policy π^* (e.g., see Puterman (1994)).

In a companion paper (Greenwald et al., 2005), we investigate the question of whether or not this iterative algorithm converges to correlated equilibrium policies in Markov games. Following the literature on this subject (e.g., Littman (1994, 2001); Hu and Wellman (2003)), we primarily experiment with "correlated Q-learning" (i.e., asynchronous updating), rather than "correlated value iteration" (i.e., synchronous updating).

A. Fixed Point Theorems: A Review

In this appendix, we include several preliminary definitions, theorems, and observations.[5] Most importantly, we present Kakutani's fixed point theorem. This theorem, which provides sufficient conditions for a correspondence to have a fixed point, generalizes Brouwer's fixed point theorem, which provides sufficient conditions for a function to have a fixed point.

In the main body of the paper, we apply Kakutani's fixed point theorem to prove the existence of stationary correlated equilibrium policies in Markov games. We define a correspondence, the fixed points of which are the stationary correlated equilibria in a Markov game, and we show that this correspondence satisfies Kakutani's sufficient conditions.

Definition 19 *A set $S \subseteq \mathbb{R}^m$ is **closed** if, for any sequence $\{x^n\}$ of vectors in S with $x^n \to x$, the limit x is also in S.*[6] *A set $S \subseteq \mathbb{R}^m$ is **bounded** if there exists a constant $N < \infty$ s.t. $\|x\|_2 \leq N$, for all $x \in S$. A set $S \subseteq \mathbb{R}^m$ is **compact** if it is closed and bounded. A set $S \subseteq \mathbb{R}^m$ is **convex** if for all $x, y \in S$, for all $\beta \in [0,1]$, $\beta x + (1-\beta)y \in S$.*

Definition 20 *(Border, Definition 11.1) A **correspondence** μ from X to Y, denoted $\mu : X \twoheadrightarrow Y$, is a function from X to the family of subsets of Y. A **fixed point of a correspondence** μ is a point x such that $x \in \mu(x)$.*

Definition 21 *(Border, Definition 11.5) The correspondence $\mu : X \twoheadrightarrow Y$ is said to be **closed at x** if whenever $x^n \to x$, $y^n \in \mu(x^n)$, and $y^n \to y$, then $y \in \mu(x)$. A correspondence that is closed everywhere is said to be **closed**.*

Theorem 22 (Kakutani's Fixed Point Theorem) *(Border, Corollary 15.3) If $K \subseteq \mathbb{R}^m$ is nonempty, compact, and convex, and if the correspondence $\lambda : K \twoheadrightarrow K$ is closed with nonempty, compact, convex values, then λ has a fixed point.*

5. We refer the reader who is seeking more details about this preliminary material to the classic text by Border (1985).

6. The notation $x^n \to x$ is an abbreviation for $\lim_{n\to\infty} x^n = x$.

Definition 23 *A function $f : X \to Y$ is **continuous at x** if, whenever $x^n \to x$ and $y^n = f(x^n)$, it is also the case that $y^n \to y$ and $y = f(x)$. A function that is continuous everywhere is said to be **continuous**.*

Lemma 24 *A continuous function $f : X \to Y$ viewed as a correspondence is closed with nonempty, compact, convex values.*

Proof A singleton set has nonempty, compact, convex values. Given a sequence x^n and a sequence y^n where $f(x^n) = y^n$ for all n, if $x^n \to x \in X$, then $y^n \to y = f(x) \in Y$, by the definition of a continuous function. Therefore, f is a closed correspondence. ■

Corollary 25 (Brouwer's Fixed Point Theorem) *If $K \subseteq \mathbb{R}^m$ is nonempty, compact, and convex, and if the function $f : K \to K$ is continuous, then f has a fixed point: i.e., there exists $k \in K$ such that $k = f(k)$.*

Proof Brouwer's fixed point theorem follows directly from Kakutani's fixed point theorem, by viewing a function as a correspondence with singleton values, and applying Lemma 24. ■

In Figure 1(a), the correspondence $\nu(x)$ defined below is plotted.

$$\nu(x) = \begin{cases} 2 & \text{if } x \in [0,1) \\ [2,4] & \text{if } x = 1 \\ 4 & \text{if } x \in (1,3) \\ [1,4] & \text{if } x = 3 \\ 1 & \text{if } x \in (3,6) \\ [1,3] & \text{if } x \in [6,8] \\ 1 & \text{if } x \in (8,10] \end{cases} \tag{21}$$

Note that $\nu(x)$ is closed with nonempty, compact, convex values. Therefore, $\nu(x)$ has a fixed point: i.e, $3 \in \nu(3)$. In contrast, the correspondence plotted in Figure 1(b) is not closed at 6, whereas the correspondence plotted in Figure 1(c) is empty in the interval $(4, 6)$, and the correspondence plotted in Figure 1(d) is not convex-valued at 6. (All closed correspondences mapping from compact sets to compact sets are compact-valued; hence, no closed correspondence with a value that is not compact can be depicted.) None of these latter three correspondences has a fixed point. If even one of Kakutani's sufficient conditions is violated, the existence of a fixed point cannot be guaranteed.

Definition 26 *The **graph** of a correspondence $\mu : X \twoheadrightarrow Y$ is the set $Gr\ \mu = \{x \times \mu(x) \mid x \in X\} \subseteq X \times Y$.*

Observe that

$$\text{Gr } \nu = [0,1)\times\{2\}\cup\{1\}\times[2,4]\cup(1,3)\times\{4\}\cup\{3\}\times[1,4]\cup(3,6)\times\{1\}\cup[6,8]\times[1,3]\cup(8,10]\times\{1\} \tag{22}$$

Remark 27 *A correspondence is closed if and only if its graph is a closed set.*

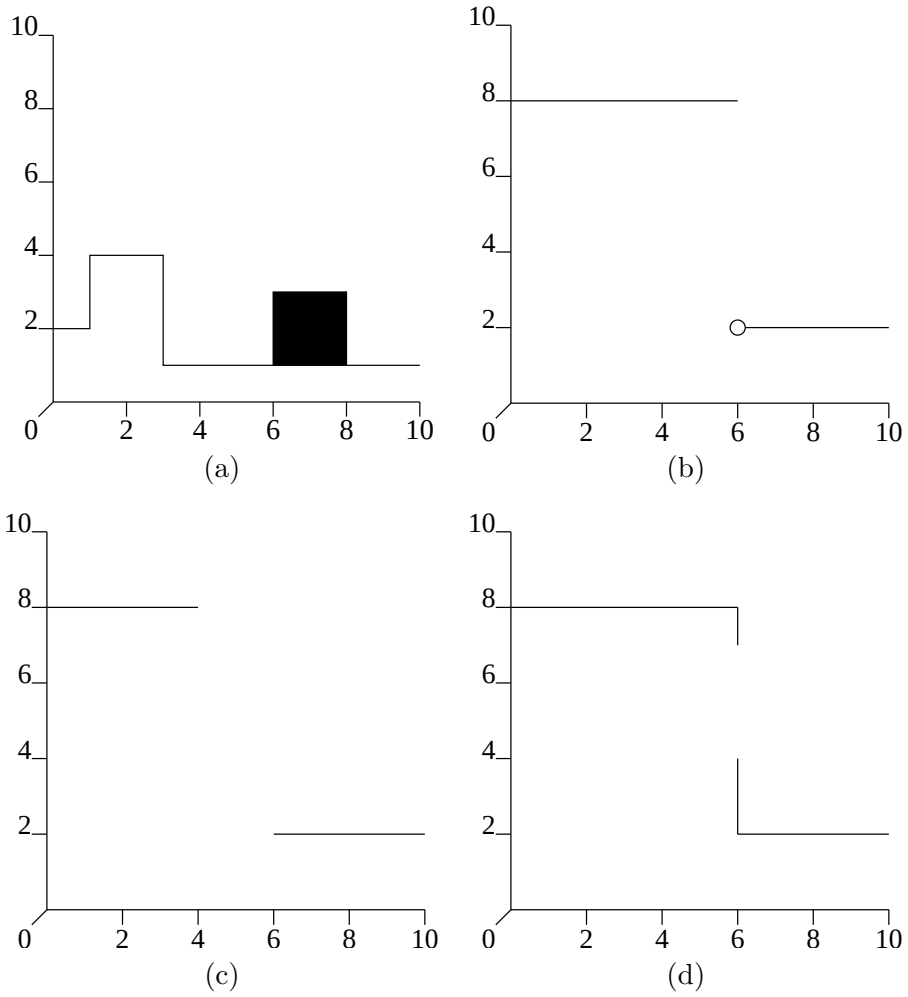

Figure 1: (a) The correspondence $\nu(x)$ defined in Equation 21. Note that $\nu(x)$ is closed with nonempty, compact, and convex values. Therefore, $\nu(x)$ has a fixed point: i.e., $3 \in \nu(3)$. (b) A correspondence that is not closed. (c) A closed correspondence with empty values. (d) A closed correspondence with a value that is not convex.

The **intersection** of two correspondences $\lambda : X \rightarrow\!\!\!\!\rightarrow Y$ and $\mu : X \rightarrow\!\!\!\!\rightarrow Y$ is defined as:

$$(\lambda \cap \mu)(x) = \lambda(x) \cap \mu(x) \tag{23}$$

Lemma 28 *If $\lambda : X \rightarrow\!\!\!\!\rightarrow Y$ is closed and $\mu : X \rightarrow\!\!\!\!\rightarrow Y$ is closed, then $\lambda \cap \mu$ is closed.*

Proof Gr $(\lambda \cap \mu) =$ (Gr λ)$\cap$(Gr μ). If Gr λ and Gr μ are closed sets, then so is Gr $(\lambda \cap \mu)$. ■

B. Proof of Theorem 13

Definition 29 *The **product** of two correspondences $\gamma : X \twoheadrightarrow Y$ and $\mu : X \twoheadrightarrow Z$ is a correspondence $(\gamma \times \mu) : X \twoheadrightarrow Y \times Z$* s.t. *$(\gamma \times \mu)(x) = \gamma(x) \times \mu(x)$.*

Lemma 30 *(Border, Lemma 11.25.c) If λ and μ are closed, then $\lambda \times \mu$ is closed.*

Lemma 31 *If λ and μ are closed with nonempty, compact, convex values, then $\lambda \times \mu$ is closed with nonempty, compact, convex values.*

Proof The product of two nonempty sets is nonempty; the product of two compact sets is compact; the product of two convex sets is convex. ■

Definition 32 *The **composition** of two correspondences $\gamma : Y \twoheadrightarrow Z$ and $\mu : X \twoheadrightarrow Y$ is a correspondence $(\gamma \circ \mu) : X \twoheadrightarrow Z$* s.t. *$(\gamma \circ \mu)(x) = \{z \in Z | \exists y \in Y$* s.t. *$z \in \gamma(y)$ & $y \in \mu(x)\}$.*

Lemma 33 *Given $X \subseteq \mathbb{R}^r$, $Y \subseteq \mathbb{R}^s$, and $Z \subseteq \mathbb{R}^t$, if $\mu : Y \twoheadrightarrow Z$ is closed with nonempty, compact, convex values, and if $f : X \to Y$ is continuous, then $\mu \circ f$ is closed[7] with nonempty, compact, convex values.*

Proof Any value of $\mu \circ f$ is also a value of μ, so the values of $\mu \circ f$ are nonempty, compact, and convex. To show that $\mu \circ f$ is closed, we assume: for any pair of sequences x^n and z^n, $x^n \to x \in X$, $z^n \to z \in Z$, and $z^n \in \mu(f(x^n))$; and we show: $z \in \mu(f(x))$. Define $y^n = f(x^n)$ so that $z^n \in \mu(y^n)$. Since f is continuous, $y^n \to y$ and $y = f(x)$. Therefore, $y^n \to y$, $z^n \to z$, and $z \in \mu(y^n)$, so that $z \in \mu(y) = \mu(f(x))$, because μ is closed. ■

Definition 34 *Let $X \subseteq \mathbb{R}^r$, $Y \subseteq \mathbb{R}^s$, and $Z \subseteq \mathbb{R}^t$.*

*A function $f : X \to Z$ is **affine** if there exists a matrix $A \in R^{r \times t}$ and a vector $c \in \mathbb{R}^t$ such that:*

$$f_k(x) = c_j + \sum_{i=1}^{r} A_{i,j} x_j \tag{24}$$

*A function $g : X \times Y \to Z$ is **bilinear** if there exists a tensor $B \in R^{r \times s \times t}$ such that:*

$$g_k(x, y) = \sum_{i=1}^{r} \sum_{j=1}^{s} B_{i,j,k} x_i y_j \tag{25}$$

Fact 35 *Affine and bilinear functions are continuous.*

Proof (of Theorem 13): The function $V_{\mathcal{Q} \times \Pi}$ is bilinear, and by Fact 35 continuous. The function $Q_{\mathcal{V}}$ is affine, and by Fact 35 continuous. Thus, $Q_{\mathcal{V}} \circ V_{\mathcal{Q} \times \Pi}$ is a continuous[8] function.

7. It is not the case that for all closed correspondences $\lambda : X \twoheadrightarrow Y$, $\mu \circ \lambda$ is closed, unless Y is compact. See Border Lemma 11.23.c.
8. In fact, $\mathcal{Q}_{\mathcal{V}} \circ V_{\mathcal{Q} \times \Pi}$ is a bilinear function.

Now the identity function I is continuous, and by Lemma 24, it can be viewed a closed correspondence with nonempty, compact, convex values. The correspondence $\pi_{\mathcal{Q}}$ is closed, with nonempty, compact, convex values, by assumption. Therefore, by Lemmas 30 and 31, the correspondence $I \times \pi_{\mathcal{Q}}$ is closed with nonempty, compact, convex values. Moreover, by Lemma 33, the correspondence $\rho = (I \times \pi_{\mathcal{Q}}) \circ \mathcal{Q}_{\mathcal{V}} \circ V_{\mathcal{Q}\times\Pi}$ is closed with nonempty, compact, convex values. Finally, since the domain of this correspondence, namely $\mathcal{Q}\times\Pi$, is nonempty, compact, and convex, by Kakutani's Fixed Point Theorem, ρ has a fixed point. ∎

C. Proof of Theorem 7

In this appendix, we establish the *one-shot deviation* principle for correlated equilibrium in Markov games. Recall that the definition of correlated equilibrium permits arbitrarily complex deviations on the part of an agent: for example, deviations could be nonstationary. The one-shot deviation principle shows that it suffices to consider one-shot deviations.

Imagine a Markov game with a referee sending recommendations according to a stationary policy π. By Remark 6, if an agent can gain by a one-shot deviation, then π is not a correlated equilibrium. The one-shot deviation principle states that the converse is also true: *if an agent cannot gain by a one-shot deviation, then π is a correlated equilibrium.*

Our proof of the one-shot deviation principle is straightforward: if all of the other agents but agent i play according to the referee's policy π, then from the point of view of agent i, its environment is a Markov decision process. Hence, the one-shot deviation principle for Markov decision processes (Blackwell, 1965) establishes Theorem 7.

C.1 Markov Decision Processes: A Review

A Markov decision process is a special case of a Markov game in which there is exactly one agent.

Definition 36 *A (finite) discounted **Markov decision process** (MDP) is a tuple $\mathcal{MDP}_\gamma = \langle S, A, P, R\rangle$ in which*

- *S is a finite set of states*
- *A is the agent's finite set of pure actions*[9]
- *P is a system of transition probabilities: i.e., for all $s \in S$, $a \in A$, $P[s' \mid s, a] \geq 0$ and $\sum_{s'\in S} P[s' \mid s, a] = 1$; we interpret $P[s' \mid s, a]$ as the probability that the next state is s' given that the current state is s and the current action is a*
- *$R : S \times A \times S \to [\alpha, \beta]$, where $R(s, a, s') \in [\alpha, \beta]$ the agent's reward for taking action a at state s and transitioning to state s'*
- *$\gamma \in [0, 1)$ is a discount factor*

Given $\mathcal{MDP}_\gamma$, a deterministic *stationary* policy is a function $\sigma : S \to A$.

9. In this appendix, to simplify notation, we assume the agent's actions sets are identical at all states of the MDP. Similarly, we assume the agents' actions sets are identical at all states of the Markov game.

Definition 37 *Given $\mathcal{MDP}_\gamma$ and stationary policy σ, the **value function** $V^\sigma : S \to \mathbb{R}$ and the **Q-value function** $Q^\sigma : S \times A \to \mathbb{R}$ are the unique functions satisfying the equations:*

$$V^\sigma(s) = Q^\sigma(s, \sigma(s)) \tag{26}$$

$$Q^\sigma(s, a) = \sum_{s' \in S} P[s' \mid s, a](R(s, a, s') + \gamma V^\sigma(s')) \tag{27}$$

Definition 38 *Given $\mathcal{MDP}_\gamma$, a stationary policy $\sigma : S \to A$ is **optimal** if, for all $s \in S$, for all $a \in A$,*

$$V^\sigma(s) = \max_{a \in A} Q^\sigma(s, a) \tag{28}$$

Equivalently, for all $s \in S$, for all $a \in A$,

$$Q^\sigma(s, \sigma(s)) \geq Q^\sigma(s, a) \tag{29}$$

Although it is not necessary to restrict the agent's policies to be stationary, it is sufficient:

Theorem 39 (Bellman, 1957) *For all finite discounted Markov decision processes, there exists an optimal, deterministic, stationary policy.*

C.2 Markov Games as Markov Decision Processes

Imagine a Markov game from the point of view of agent i. If the referee's policy π is stationary, and if all the other agents except perhaps agent i follow the referee's advice, then from the point of view of agent i, the environment is an MDP. In fact, given a stationary policy π for the referee, any Markov game can be expressed as an MDP for agent i, such that the information available to the agent in both the Markov game and the MDP is identical. We now demonstrate this fact by construction.

From agent i's perspective, a stage in a Markov game with a referee proceeds as follows:

Process 1: Markov Game

1. Agent i observes the state.
2. Agent i receives advice from the referee.
3. Agent i selects an action.
4. Agent i observes the actions of the other agents.
5. Agent i receives its reward.

In a Markov game, the referee sends its advice to all agents simultaneously. In the MDP, however, from the perspective of agent i, it is equivalent to imagine that first the referee sends its recommendation to agent i and agent i selects its action, and then the referee sends its recommendations to the other agents and the other agents select their actions. Thus, we can break down the stages in a different fashion:

Process 2: Markov Decision Process

1. Agent i observes the actions of the other agents during the previous stage of the Markov game.
2. Agent i observes the reward it received during the previous stage of the Markov game.
3. Agent i observes the current state of the Markov game.
4. Agent i receives advice from the referee.
5. Agent i selects an action.

In order to represent a Markov game $\Gamma_\gamma = \langle N, S, A, P, R\rangle$ from the point of view of agent i as an MDP, we include in the MDP's state, from the Markov game, the previous actions of the other agents, the current state, and the referee's advice.

Formally, we define $\mathcal{MDP}_{\gamma'} = \langle S', A', P', R'\rangle$ as follows:

1. $S' = \{(a_{-i}, s, a_i) \in A_{-i} \times S \times A_i \mid \pi_s(a_i) > 0\}$
2. $A' = A_i$
3. $P'[(a''_{-i}, s'', a''_i) \mid (a_{-i}, s, a_i), a'_i] = \pi_s(a''_{-i} \mid a_i)P[s'' \mid s, (a'_i, a''_{-i})]\pi_{s''}(a''_i)$
4. $R'((a_{-i}, s, a_i), a''_i, (a'_{-i}, s', a'_i)) = R(s, (a''_i, a'_{-i}))$
5. $\gamma' = \gamma$

Observe that, given the history of past states and past actions in this MDP, agent i knows, in the Markov game, the past actions of the other agents, the history of past states, the referee's recommended actions for agent i, and agent i's actions. Thus, any policy that agent i can implement in the Markov game can also be implemented in the MDP. More importantly, *the expected utility of any policy in the Markov game is equal to the expected utility of its implementation in the MDP.*

Before proving the one-shot deviation principle for Markov games, we first elaborate on this key point. Specifically, we describe the mathematical relationship between the agent's expected utility of following the advice of the referee in the Markov game and following the advice of the referee in the MDP. Suppose that π is the referee's policy in the Markov game, and that σ is the agent's policy in the MDP that dictates following the advice of the referee: i.e., for all $(a_{-i}, s, a_i) \in S'$, $\sigma(a_{-i}, s, a_i) = a_i$.

Define the following random variables:

1. $\mathbf{s}_t$, the state at time t;
2. $\mathbf{a}_{i,t}$, the action recommended to agent i at time t;
3. $\mathbf{a}_{-i,t}$, the actions recommended to the agents other than i at time t;
4. $\mathbf{p}_{i,t}$, the action actually played by agent i at time t;
5. $\mathbf{o}_t$, the utility observed at time t in Process 2;
6. $\mathbf{r}_t$, the utility received at time t in Process 1;
7. θ, the assumption that all agents follow the advice of the referee after time 1, and that all agents but i follow the advice of the referee at time 1.

Looking back at the Markov game and the MDP, the observed reward in Process 2 is the reward received during the previous round in Process 1: i.e., $\mathbf{r}_t = \mathbf{o}_{t+1}$. Therefore, for all $(a_{-i}, s, a_i) \in S'$ (i.e., $\pi_s(a_i) > 0$), for all $a'_i \in A_i$:

$$Q^\sigma((a_{-i}, s, a_i), a'_i) = \mathbb{E}\left[\sum_{t=2}^{\infty} \mathbf{o}_t \gamma^{t-2} \mid \mathbf{s}_1 = s, \mathbf{a}_{i,1} = a_i, \mathbf{p}_{i,1} = a'_i, \theta\right] \tag{30}$$

$$= \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbf{r}_t \gamma^{t-1} \mid \mathbf{s}_1 = s, \mathbf{a}_{i,1} = a_i, \mathbf{p}_{i,1} = a'_i, \theta\right] \tag{31}$$

$$= \sum_{a'_{-i} \in A_{-i}} \Pr[\mathbf{a}_{-i,1} = a'_{-i} | \mathbf{s}_1 = s, \mathbf{a}_{i,1} = a_i, \mathbf{p}_{i,1} = a'_i, \theta]$$

$$\times \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbf{r}_t \gamma^{t-1} \mid \mathbf{a}_{-i,1} = a'_{-i}, \mathbf{s}_1 = s, \mathbf{a}_{i,1} = a_i, \mathbf{p}_{i,1} = a'_i, \theta\right] \tag{32}$$

$$= \sum_{a'_{-i} \in A_{-i}} \pi_s(a'_{-i} \mid a_i) Q_i^\pi(s, (a'_{-i}, a'_i)) \tag{33}$$

C.3 One-Shot Deviation Principle

Proof (of Theorem 7): Given a Markov game $\Gamma_\gamma = \langle N, S, A, P, R \rangle$ and a policy π for the referee, construct $\mathcal{MDP}_{\gamma'} = \langle S', A', P', R' \rangle$ as above. Assume agent i, in this MDP, always follows the advice of the referee: for all $a_{-i} \in A_{-i}$, $s \in S$, $a_i \in A_i$, $\sigma(a_{-i}, s, a_i) = a_i$.

By the one-shot deviation principle for MDPs, σ is an optimal policy in agent i's MDP if and only if for all $a_{-i} \in A_{-i}$, for all $s \in S$, for all $a_i \in A_i$ with $\pi_s(a_i) > 0$, for all $a'_i \in A_i$,

$$Q^\sigma((a_{-i}, s, a_i), a_i) \geq Q^\sigma((a_{-i}, s, a_i), a'_i) \tag{34}$$

By Equation 33, the above condition holds if and only if for all $s \in S$, for all $a_i \in A_i$ with $\pi_s(a_i) > 0$, for all $a'_i \in A_i$,

$$\sum_{a''_{-i} \in A_{-i}} \pi_s(a''_{-i} \mid a_i) Q_i^\pi(s, (a''_{-i}, a_i)) \geq \sum_{a''_{-i} \in A_{-i}} \pi_s(a''_{-i} \mid a_i) Q_i^\pi(s, (a''_{-i}, a'_i)) \tag{35}$$

that is, agent i does not stand to gain from any one-shot deviations in the Markov game.

Finally, imagine n such MDPs, one per agent. By construction, the referee's policy π is a correlated equilibrium in the Markov game if and only if σ is an optimal policy in agent i's MDP, for all agents i. We conclude that π is a correlated equilibrium if and only if no agent can gain from any one-shot deviations in the Markov game. ∎

Acknowledgments

The authors are grateful to Michael Littman for inspiring discussions. We also thank Keith Hall, Dinah Rosenberg, and Roberto Serrano for comments on an earlier draft of this paper. This research was supported by NSF Career Grant #IIS-0133689.

References

R. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

R. Bellman. *Dynamic Programming*. Princeton University Press, New Jersey, 1957.

D. Blackwell. Discounted dynamic programming. *Annals of Mathematical Statistics*, 36: 226–235, 1965.

K. Border. *Fixed Point Theorems with Applications to Economics and Game Theory*. Cambridge University Press, Cambridge, 1985.

Vincent Conitzer and Tuomas Sandholm. Complexity results about nash equilibria. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 765–771, August 2003.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer Verlag, New York, 1996.

A.M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science in Hiroshima University*, 28:89–93, 1964.

Francoise Forges. An approach to communication equilibria. *Econometrica*, 54(6):1375–1385, 1986.

Itzhak Gilboa and Eitan Zemel. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1:80–93, 1989.

A. Greenwald, K. Hall, and M. Zinkevich. Correlated Q-learning. Technical Report 8, Brown University, Department of Computer Science, June 2005.

J. Hu and M. Wellman. Nash Q-learning for general-sum stochastic games. *Machine Learning Research*, 4:1039–1069, 2003.

M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163, July 1994.

M. Littman. Friend or foe Q-learning in general-sum Markov games. In *Proceedings of Eighteenth International Conference on Machine Learning*, pages 322–328, June 2001.

J. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

A. Neyman and S. Sorin, editors. *Stochastic Games and Applications*. Kluwer Academic Publishers, NATO ASI Series, 2003.

M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, 1994.

M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994.

L.S. Shapley. Stochastic games. In H. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*, volume II, pages 80–86. Princeton University Press, 1953a.

L.S. Shapley. A value for n-person games. In H. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games*, volume II, pages 307–317. Princeton University Press, 1953b.

K. Shell. General equilibrium: The new palgrave. In J. Eatwell, M. Milgate, and P. Newman, editors, *Sunspot Equilibrium*, pages 274–280. Macmillan, New York, 1989.

Eilon Solan and Nicolas Vieille. Correlated equilibrium in stochastic games. *Games and Economic Behavior*, 38(2):362–399, 2002.

J. von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, 1989.