Cyclic Equilibria in Markov Games

Martin Zinkevich and Amy Greenwald
Department of Computer Science
Brown University
Providence, RI 02912
{maz,amy}@cs.brown.edu

Michael L. Littman
Department of Computer Science
Rutgers, The State University of NJ
Piscataway, NJ 08854–8019
mlittman@cs.rutgers.edu

Abstract

Although variants of value iteration have been proposed for finding Nash or correlated equilibria in general-sum Markov games, these variants have not been shown to be effective in general. In this paper, we demonstrate by construction that existing variants of value iteration cannot find stationary equilibrium policies in arbitrary general-sum Markov games. Instead, we propose an alternative interpretation of the output of value iteration based on a new (non-stationary) equilibrium concept that we call "cyclic equilibria." We prove that value iteration identifies cyclic equilibria in a class of games in which it fails to find stationary equilibria. We also demonstrate empirically that value iteration finds cyclic equilibria in nearly all examples drawn from a random distribution of Markov games.

1 Introduction

Value iteration (Bellman, 1957) has proven its worth in a variety of sequential-decision-making settings, most significantly single-agent environments (Puterman, 1994), team games, and two-player zero-sum games (Shapley, 1953). In value iteration for Markov decision processes and team Markov games, the value of a state is defined to be the maximum over all actions of the value of the combination of the state and action (or *Q value*). In zero-sum environments, the max operator becomes a minimax over joint actions of the two players. Learning algorithms based on this update have been shown to compute equilibria in both model-based scenarios (Brafman & Tennenholtz, 2002) and Q-learning-like model-free scenarios (Littman & Szepesvári, 1996).

The theoretical and empirical success of such algorithms has led researchers to apply the same approach in general-sum games, in spite of exceedingly weak guarantees of convergence (Hu & Wellman, 1998; Greenwald & Hall, 2003). Here, value-update rules based on select Nash or correlated equilibria have been evaluated empirically and have been shown to perform reasonably in some settings. None has been identified that computes equilibria in general, however, leaving open the question of whether such an update rule is even possible.

Our main theoretical result is that an entire class of value-iteration update rules, including all those mentioned above, can be excluded from consideration for computing stationary equilibria in general-sum Markov games. Briefly, existing value-iteration algorithms compute Q values as an intermediate result, then derive policies from these Q values. We demonstrate a class of games in which Q values, even those corresponding to an equilibrium policy, contain insufficient information for reconstructing an equilibrium policy.

Faced with the impossibility of developing algorithms along the lines of traditional value iteration that find stationary equilibria, we suggest an alternative equilibrium concept—cyclic equilibria. A cyclic equilibrium is a kind of non-stationary joint policy that satisfies the standard conditions for equilibria (no incentive to deviate unilaterally). However, unlike conditional non-stationary policies such as tit-for-tat and finite-state strategies based on the "folk theorem" (Osborne & Rubinstein, 1994), cyclic equilibria cycle sequentially through a set of stationary policies.

We present two positive results concerning cyclic equilibria. First, we consider the class of two-player two-state two-action games used to show that Q values cannot reconstruct all stationary equilibrium. Section 4.1 shows that value iteration finds cyclic equilibria for all games in this class. Second, Section 5 describes empirical results on a more general set of games. We find that on a significant fraction of these games, value iteration updates fail to converge. In contrast, value iteration finds cyclic equilibria for nearly all the games.

The success of value iteration in finding cyclic equilibria suggests this generalized solution concept could be useful for constructing robust multiagent learning algorithms.

2 An Impossibility Result for Q Values

In this section, we consider a subclass of Markov games in which transitions are deterministic and are controlled by one player at a time. We show that this class includes games that have no deterministic equilibrium policies. For this class of games, we present (proofs available in an extended technical report) two theorems. The first, a negative result, states that the Q values used in existing value-iteration algorithms are insufficient for deriving equilibrium policies. The second, presented in Section 4.1, is a positive result that states that value iteration does converge to cyclic equilibrium policies in this class of games.

2.1 Preliminary Definitions

Given a finite set X, define $\Delta(X)$ to be the set of all probability distributions over X.

Definition 1 *A* ***Markov game*** $\Gamma = [S, N, \mathbf{A}, T, R, \gamma]$ *is a set of states S, a set of players $N = \{1, \ldots, n\}$, a set of actions for each player in each state $\{A_{i,s}\}_{s \in S, i \in N}$ (where we represent the set of all state-action pairs as $\mathbf{A} \equiv \bigcup_{s \in S} \left(\{s\} \times \prod_{i \in N} A_{i,s}\right)$, a transition function $T : \mathbf{A} \to \Delta(S)$, a reward function $R : \mathbf{A} \to \mathbf{R}^n$, and a discount factor γ.*

Given a Markov game Γ, let $A_s = \prod_{i \in N} A_{i,s}$. A **stationary policy** is a set of distributions $\{\pi(s) : s \in S\}$, where for all $s \in S$, $\pi(s) \in \Delta(A_s)$. Given a stationary policy π, define $V^{\pi,\Gamma} : S \to \mathbf{R}^n$ and $Q^{\pi,\Gamma} : \mathbf{A} \to \mathbf{R}^n$ to be the unique pair of functions satisfying the following system of equations: for all $i \in N$, for all $(s, a) \in \mathbf{A}$,

$$V_i^{\pi,\Gamma}(s) = \sum_{a \in A_s} \pi(s)(a) Q_i^{\pi,\Gamma}(s, a), \tag{1}$$

$$Q_i^{\pi,\Gamma}(s, a) = R_i(s, a) + \gamma \sum_{s' \in S} T(s, a)(s') V_i^{\pi,\Gamma}(s'). \tag{2}$$

A **deterministic** Markov game is a Markov game Γ where the transition function is deterministic: $T : \mathbf{A} \to S$. A **turn-taking game** is a Markov game Γ where in every state, only one player has a choice of action. Formally, for all $s \in S$, there exists a player $i \in N$ such that for all other players $j \in N \backslash i$, $|A_{j,s}| = 1$.

2.2 A Negative Result for Stationary Equilibria

A **NoSDE** (pronounced "nasty") game is a deterministic turn-taking Markov game Γ with two players, two states, no more than two actions for either player in either state, and no deterministic stationary equilibrium policy. That the set of NoSDE games is non-empty is demonstrated by the game depicted in Figure 1. This game has no deterministic stationary equilibrium policy: If Player 1 sends, Player 2 prefers to send; but, if Player 2 sends, Player 1 prefers to keep; and, if Player 1 keeps, Player 2 prefers to keep; but, if Player 2 keeps, Player 1 prefers to send. No deterministic policy is an equilibrium because one player will always have an incentive to change policies.

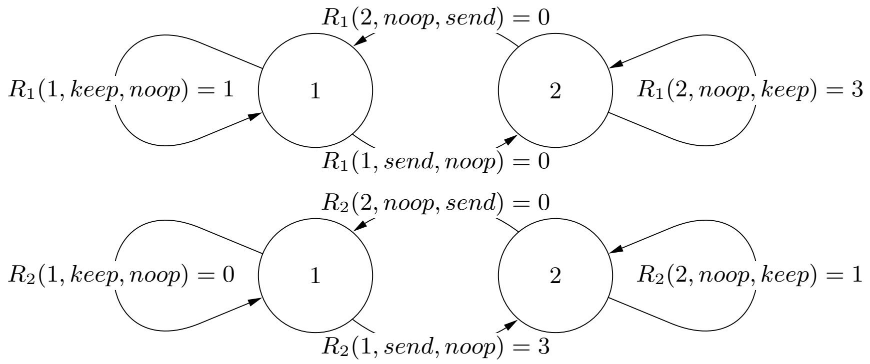

Figure 1: An example of a NoSDE game. Here, $S = \{1, 2\}$, $A_{1,1} = A_{2,2} = \{keep, send\}$, $A_{1,2} = A_{2,1} = \{noop\}$, $T(1, keep, noop) = 1$, $T(1, send, noop) = 2$, $T(2, noop, keep) = 2$, $T(2, noop, send) = 1$, and $\gamma = 3/4$. In the unique stationary equilibrium, Player 1 sends with probability $2/3$ and Player 2 sends with probability $5/12$.

Lemma 1 *Every NoSDE game has a* unique *stationary equilibrium policy.*[1]

It is well known that, in general Markov games, random policies are sometimes needed to achieve an equilibrium. This fact can be demonstrated simply by a game with one state where the utilities correspond to a bimatrix game with no deterministic equilibria (penny matching, say). Random actions in these games are sometimes linked with strategies that use "faking" or "bluffing" to avoid being exploited. That NoSDE games exist is surprising, in that randomness is needed even though actions are always taken with complete information about the other player's choice and the state of the game. However, the next result is even more startling. Current value-iteration algorithms attempt to find the Q values of a game with the goal of using these values to find a stationary equilibrium of the game. The main theorem of this paper states that it is not possible to derive a policy from the Q values for NoSDE games, and therefore in general Markov games.

Theorem 1 *For any NoSDE game $\Gamma = [S, N, \mathbf{A}, T, R]$ with a unique equilibrium policy π, there exists another NoSDE game $\Gamma' = [S, N, \mathbf{A}, T, R']$ with its own unique equilibrium policy π' such that $Q^{\pi,\Gamma} = Q^{\pi',\Gamma'}$ but $\pi \neq \pi'$ and $V^{\pi,\Gamma} \neq V^{\pi',\Gamma'}$.*

This result establishes that computing or learning Q values is insufficient to compute a stationary equilibrium of a game.[2] In this paper we suggest an alternative, where we still

[1] The policy is both a correlated equilibrium and a Nash equilibrium.

[2] Although maintaining Q values *and* state values and deriving policies from both sets of functions might circumvent this problem, we are not aware of existing value-iteration algorithms or learning algorithms that do so. This observation presents a possible avenue of research not followed in this paper.

do value iteration in the same way, but we extract a *cyclic equilibrium* from the sequence of values instead of a stationary one.

3 A New Goal: Cyclic Equilibria

A **cyclic policy** is a finite sequence of stationary policies $\pi = \{\pi_1, \ldots, \pi_k\}$. Associated with π is a sequence of value functions $\{V^{\pi,\Gamma,j}\}$ and Q-value functions $\{Q^{\pi,\Gamma,j}\}$ such that

$$V_i^{\pi,\Gamma,j}(s) = \sum_{a \in A_s} \pi_j(s)(a)\, Q_i^{\pi,\Gamma,j}(s,a) \quad \text{and} \tag{3}$$

$$Q_i^{\pi,\Gamma,j}(s,a) = R_i(s,a) + \gamma \sum_{s' \in S} T(s,a)(s')\, V_i^{\pi,\Gamma,inc_k(j)}(s') \tag{4}$$

where for all $j \in \{1, \ldots, k\}$, $inc_k(k) = 1$ and $inc_k(j) = j+1$ if $j < k$.

Definition 2 *Given a Markov game Γ, a* ***cyclic correlated equilibrium*** *is a cyclic policy π, where for all $j \in \{1, \ldots, k\}$, for all $i \in N$, for all $s \in S$, for all $a_i, a_i' \in A_{i,s}$:*

$$\sum_{a_{-i} \in A_{-i,s}} \pi_j(s)(a_i, a_{-i})\, Q_i^{\pi,\Gamma,j}(s, a_i, a_{-i}) \geq \sum_{(a_i, a_{-i}) \in A_s} \pi_j(s)(a_i, a_{-i})\, Q^{\pi,\Gamma,j}(s, a_i', a_{-i}). \tag{5}$$

Here, a_{-i} denotes a joint action for all players except i. A similar definition can be constructed for Nash equilibria by insisting that all policies $\pi_j(s)$ are product distributions. In Definition 2, we imagine that action choices are moderated by a referee with a clock that indicates the current stage j of the cycle. At each stage, a typical correlated equilibrium is executed, meaning that the referee chooses a joint action a from $\pi_j(s)$, tells each agent its part of that joint action, and no agent can improve its value by eschewing the referee's advice. If no agent can improve its value by more than ϵ at any stage, we say π is an **ϵ-correlated cyclic equilibrium**.

A **stationary correlated equilibrium** is a cyclic correlated equilibrium with $k = 1$. In the next section, we show how value iteration can be used to derive cyclic correlated equilibria.

4 Value Iteration in General-Sum Markov Games

For a game Γ, define $\mathcal{Q}_\Gamma = (\mathbf{R}^n)^{\mathbf{A}}$ to be the set of all state-action (Q) value functions, $\mathcal{V}_\Gamma = (\mathbf{R}^n)^S$ to be the set of all value functions, and Π_Γ to be the set of all stationary policies. Traditionally, value iteration can be broken down into estimating a Q value based upon a value function, selecting a policy π given the Q values, and deriving a value function based upon π and the Q-value functions. Whereas the first and the last step are fairly straightforward, the step in the middle is quite tricky. A pair $(\pi, Q) \in \Pi_\Gamma \times \mathcal{Q}_\Gamma$ **agree** (see Equation 5) if, for all $s \in S$, $i \in N$, $a_i, a_i' \in A_{i,s}$:

$$\sum_{a_{-i} \in A_{-i,s}} \pi(s)(a_i, a_{-i})\, Q_i(s, a_i, a_{-i}) \geq \sum_{(a_i, a_{-i}) \in A_s} \pi(s)(a_i, a_{-i})\, Q(s, a_i', a_{-i}). \tag{6}$$

Essentially, Q and π agree if π is a best response for each player given payoffs Q. An **equilibrium-selection rule** is a function $f : \mathcal{Q}_\Gamma \to \Pi_\Gamma$ such that for all $Q \in \mathcal{Q}_\Gamma$, $(f(Q), Q)$ agree. In essence, these rules update values assuming an equilibrium policy for a one-stage game with $Q(s,a)$ providing the terminal rewards. Examples of equilibrium-selection rules are best-Nash, utilitarian-CE, dictatorial-CE, plutocratic-CE, and egalitarian-CE (Greenwald & Hall, 2003). (Utilitarian-CE, which we return to later, selects the correlated equilibrium in which total of the payoffs is maximized.) Foe-VI and Friend-VI (Littman, 2001) do not fit into our formalism, but it can be proven that in NoSDE games they converge to deterministic policies that are neither stationary nor cyclic equilibria. Define $d_\Gamma : \mathcal{V}_\Gamma \times \mathcal{V}_\Gamma \rightarrow \mathbf{R}$ to be a distance metric over value functions, such that

$$d_\Gamma(V, V') = \max_{s \in S, i \in N} |V_i(s) - V'_i(s)|. \tag{7}$$

Using our notation, the value-iteration algorithm for general-sum Markov games can be described as follows.

Algorithm 1: ValueIteration(game Γ, $V^0 \in \mathcal{V}_\Gamma$, $f \in \mathbf{F}_\Gamma$, Integer T)
For $t := 1$ to T:

1. $\forall s \in S, a \in A, Q^t(s,a) := R(s,a) + \gamma \sum_{s' \in S} T(s,a)(s')\ V^{t-1}(s')$.
2. $\pi^t = f(Q^t)$.
3. $\forall s \in S, V^t(s) = \sum_{a \in A_s} \pi^t(s)(a)\ Q^t(s,a)$.

Return $\{Q^1, \ldots, Q^T\}$, $\{\pi^1, \ldots, \pi^T\}$, $\{V^1, \ldots, V^T\}$.

If a stationary equilibrium is sought, the final policy is returned.

Algorithm 2: GetStrategy(game Γ, $V^0 \in \mathcal{V}_\Gamma$, $f \in \mathbf{F}_\Gamma$, Integer T)

1. Run $(Q^1 \ldots Q^T, \pi^1 \ldots \pi^T, V^1 \ldots V^T) = \text{ValueIteration}(\Gamma, V^0, f, T)$.
2. Return π^T.

For cyclic equilibria, we have a variety of options for how many past stationary policies we want to consider for forming a cycle. Our approach searches for a recent value function that matches the final value function (an exact match would imply a true cycle). Ties are broken in favor of the shortest cycle length. Observe that the order of the policies is reversed to form a cyclic equilibrium.

Algorithm 3: GetCycle(game Γ, $V^0 \in V_\Gamma$, $f \in \mathbf{F}_\Gamma$, Integer T, Integer $maxCycle$)

1. If $maxCycle \geq T$, $maxCycle := T - 1$.
2. Run $(Q^1 \ldots Q^T, \pi^1 \ldots \pi^T, V^1 \ldots V^T) = \text{ValueIteration}(\Gamma, V^0, f, T)$.
3. Define $k := \text{argmin}_{t \in \{1,\ldots,maxCycle\}}\ d(V^T, V^{T-t})$.
4. For each $t \in \{1, \ldots, k\}$ set $\pi_t := \pi^{T+1-t}$.

4.1 Convergence Conditions

Fact 1 *If $d(V^T, V^{T-1}) = \epsilon$ in **GetStrategy**, then **GetStrategy** returns an $\frac{\epsilon\gamma}{1-\gamma}$-correlated equilibrium.*

Fact 2 *If **GetCycle** returns a cyclic policy of length k and $d(V^T, V^{T-k}) = \epsilon$, then **GetCycle** returns an $\frac{\epsilon\gamma}{1-\gamma^k}$-correlated cyclic equilibrium.*

Since, given V^0 and Γ, the space of value functions is bounded, *eventually* there will be two value functions in $\{V^1, \ldots, V^T\}$ that are close according to d_Γ. Therefore, the two practical (and open) questions are (1) how many iterations does it take to find an ϵ-correlated cyclic equilibrium? and (2) How large is the cyclic equilibrium that is found?

In addition to approximate convergence described above, in two-player turn-taking games, one can prove *exact convergence*. In fact, all the members of $\mathbf{F}_\Gamma$ described above can be construed as generalizations of utilitarian-CE in turn-taking games, and utilitarian-CE is proven to converge.

Theorem 2 *Given the utilitarian-CE equilibrium-selection rule f, for every NoSDE game Γ, for every $V^0 \in \mathcal{V}_\Gamma$, there exists some finite T such that* ***GetCycle***$(\Gamma, V^0, f, T, \lceil T/2 \rceil)$ *returns a cyclic correlated equilibrium.*

Theoretically, we can imagine passing infinity as a parameter to value iteration. Doing so shows the limitation of value-iteration in Markov games.

Theorem 3 *Given the utilitarian-CE equilibrium-selection rule f, for any NoSDE game Γ with unique equilibrium π, for every $V^0 \in \mathcal{V}_\Gamma$, the value-function sequence $\{V^1, V^2, \ldots\}$ returned from* ***ValueIteration***(Γ, V^0, f, ∞) *does not converge to V^π.*

Since all of the other rules specified above (except friend-VI and foe-VI) can be implemented with the utilitarian-CE equilibrium-selection rule, none of these rules will be guaranteed to converge, even in such a simple class as turn-taking games!

Theorem 4 *Given the game Γ in Figure 1 and its stationary equilibrium π, given $V_i^0(s) = 0$ for all $i \in N$, $s \in S$, then for any update rule $f \in \mathbf{F}_\Gamma$, the value-function sequence $\{V^1, V^2, \ldots\}$ returned from* ***ValueIteration***(Γ, V^0, f, ∞) *does not converge to V^π.*

5 Empirical Results

To complement the formal results of the previous sections, we ran two batteries of tests on value iteration in randomly generated games. We assessed the convergence behavior of value iteration to stationary and cyclic equilibria.

5.1 Experimental Details

Our game generator took as input the set of players N, the set of states S, and for each player i and state s, the actions $A_{i,s}$. To construct a game, for each state-joint action pair $(s, a) \in \mathbf{A}$, for each agent $i \in N$, the generator sets $R_i(s, a)$ to be an integer between 0 and 99, chosen uniformly at random. Then, it selects $T(s, a)$ to be deterministic, with the resulting state chosen uniformly at random. We used a consistent discount factor of $\gamma = 0.75$ to decrease experimental variance.

The primary dependent variable in our results was the frequency with which value iteration converged to a stationary Nash equilibrium or a cyclic Nash equilibrium (of length less than 100). To determine convergence, we first ran value iteration for 1000 steps. If $d_\Gamma(V^{1000}, V^{999}) \leq 0.0001$, then we considered value iteration to have converged to a stationary policy. If for some $k \leq 100$

$$\max_{t \in \{1,\ldots,k\}} d_\Gamma(V^{1001-t}, V^{1001-(t+k)}) \leq 0.0001, \tag{8}$$

then we considered value iteration to have converged to a cycle.[3]

To determine if a game has a deterministic equilibrium, for every deterministic policy π, we ran policy evaluation (for 1000 iterations) to estimate $V^{\pi,\Gamma}$ and $Q^{\pi,\Gamma}$, and then checked if π was an ϵ-correlated equilibrium for ϵ=0.0001.

[3] In contrast to the **GetCycle** algorithm, we are here concerned with finding a cyclic equilibrium so we check an entire cycle for convergence.

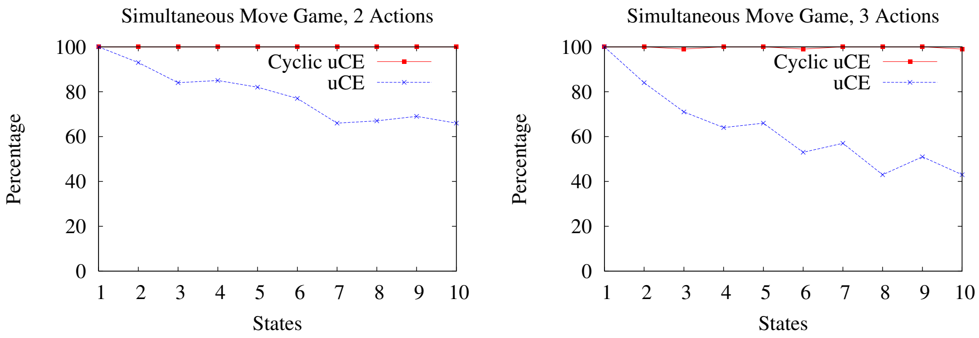

Figure 2: (Left) For each combination of states and actions, 1000 deterministic turn-taking games were generated. The graph plots the number of games where value iteration did not converge to a stationary equilibrium. (Right) Frequency of convergence on 100 randomly generated games with simultaneous actions. **Cyclic uCE** is the number of times utilitarian-CE converged to a cyclic equilibrium. **OTComb** is the number of games where any one of Friend-VI, Foe-VI, utilitarian-NE-VI, and 5 variants of correlated equilibrium-VI: dictatorial-CE-VI (First Player), dictatorial-CE-VI (Second Player), utilitarian-CE-VI, plutocratic-CE-VI, and egalitarian-VI converged to a stationary equilibrium. **OTBest** is the maximum number of games where the best fixed choice of the equilibrium-selection rule converged. **uCE** is the number of games in which utilitarian-CE-VI converged to a stationary equilibrium.

5.2 Turn-taking Games

In the first battery of tests, we considered sets of turn-taking games with x states and y actions: formally, there were x states $\{1, \ldots, x\}$. In odd-numbered states, Player 1 had y actions and Player 2 had one action: in even-numbered states, Player 1 had one action and Player 2 had y actions. We varied x from 2 to 5 and y from 2 to 10. For each setting of x and y, we generated and tested one thousand games.

Figure 2 (left) shows the number of generated games for which value iteration did *not* converge to a stationary equilibrium. We found that nearly half (48%, as many as 5% of the total set) of these non-converged games had no stationary, deterministic equilibria (they were NoSDE games). The remainder of the stationary, deterministic equilibria were simply not discovered by value iteration. We also found that value iteration converged to cycles of length 100 or less in 99.99% of the games.

5.3 Simultaneous Games

In a second set of experiments, we generated two-player Markov games where both agents have at least two actions in every state. We varied the number of states between 2 and 9, and had either 2 or 3 actions for every agent in every state.

Figure 2 (right) summarizes results for 3-action games (2-actions games were qualitatively similar, but converged more often). Note that the fraction of random games on which the algorithms converged to stationary equilibria decreases as the number of states increases. This result holds because the larger the game, the larger the chance that value iteration will fall into a cycle on some subset of the states. Once again, we see that the cyclic equilibria are found much more reliably than stationary equilibria by value-iteration algorithms. For example, utilitarian-CE converges to a cyclic correlated equilibrium about 99% of the time, whereas with 10 states and 3 actions, on 26% of the games none of the techniques converge.

6 Conclusion

In this paper, we showed that value iteration, the algorithmic core of many multiagent planning reinforcement-learning algorithms, is not well behaved in Markov games. Among other impossibility results, we demonstrated that the Q-value function retains too little information for constructing optimal policies, even in 2-state, 2-action, deterministic turn-taking Markov games. In fact, there are an infinite number of such games with different Nash equilibrium value functions that have identical Q-value functions. This result holds for proposed variants of value iteration from the literature such as updating via a correlated equilibrium or a Nash equilibrium, since, in turn-taking Markov games, both rules reduce to updating via the action with the maximum value for the controlling player.

Our results paint a bleak picture for the use of value-iteration-based algorithms for computing stationary equilibria. However, in a class of games we called NoSDE games, a natural extension of value iteration converges to a limit cycle, which is in fact a cyclic (nonstationary) Nash equilibrium policy. Such cyclic equilibria can also be found reliably for randomly generated games and there is evidence that they appear in some naturally occurring problems (Tesauro & Kephart, 1999). One take-away message of our work is that nonstationary policies may hold the key to improving the robustness of computational approaches to planning and learning in general-sum games.

References

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Brafman, R. I., & Tennenholtz, M. (2002). R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, *3*, 213–231.

Greenwald, A., & Hall, K. (2003). Correlated Q-learning. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 242–249).

Hu, J., & Wellman, M. (1998). Multiagent reinforcement learning:theoretical framework and an algorithm. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 242–250). Morgan Kaufman.

Littman, M. (2001). Friend-or-foe Q-learning in general-sum games. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 322–328). Morgan Kaufmann.

Littman, M. L., & Szepesvári, C. (1996). A generalized reinforcement-learning model: Convergence and applications. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 310–318).

Osborne, M. J., & Rubinstein, A. (1994). *A Course in Game Theory*. The MIT Press.

Puterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley-Interscience.

Shapley, L. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, *39*, 1095–1100.

Tesauro, G., & Kephart, J. (1999). Pricing in agent economies using multi-agent Q-learning. *Proceedings of Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 71–86).

A Proof of Lemma 1

Lemma 2 *In a NoSDE game:*

1. *In one state, the first player has two actions and the second player has one action, and*
2. *in the other state, the second player has two actions and the first player has one action.*

Proof: Observe that if one of the players has no action set larger than one, then the game effectively becomes an MDP and therefore has a deterministic equilibrium. Since a NoSDE game is a turn-taking game, each state can only have one player with more than one action in that state. Since there are as many players as states, the result follows directly. ■

Lemma 3 *In any turn-taking game Γ with two states and two actions in each state, if there exist $s, s' \in S$ such that for all $a \in A$, $T(s, a) = s'$, then Γ is not a NoSDE game.*

Proof: Define $S = \{s_1, s_2\}$, $A_{1,1} = \{a_1, a_1'\}$, $A_{1,2} = \{noop\}$, $A_{2,1} = \{noop\}$, $A_{2,2} = \{a_2, a_2'\}$. If the game does not have this structure, then by Lemma 2, it is not a NoSDE game.

First, consider the case where $s = s'$. Without loss of generality, assume that $s = s_1$. Also, without loss of generality, assume that $R_1(s_1, a_1, noop) \geq R_1(s_1, a_1', noop)$. We will now establish that there is a deterministic Nash equilibrium where the first player selects action a_1 in state s_1.

For s_2, for each $a_2'' \in A_{2,2}$, define:

$$Q_2(s_2, noop, a_2'') = \begin{cases} \frac{R_2(s_2, noop, a_2'')}{1-\gamma} & \text{if } T(s_2, noop, a_2'') = s_2 \\ R_2(s_2, noop, a_2'') + \frac{\gamma R_2(s_1, a_1, noop)}{1-\gamma} & \text{otherwise} \end{cases} \tag{9}$$

Without loss of generality, assume that $Q_2(s_2, noop, a_2) \geq Q_2(s_2, noop, a_2')$. Then, the first player choosing a_1 and the second player choosing a_2 is a Nash equilibrium.

Secondly, consider the case where $s \neq s'$. Without loss of generality, assume that $s = s_1$ and $s' = s_2$. Without loss of generality, assume that $R_1(s_1, a_1, noop) \geq R_1(s_1, a_1', noop)$. We will now establish that there is a deterministic Nash equilibrium where the first player selects action a_1 in state s_1.

For s_2, for each $a_2'' \in A_{2,2}$, define:

$$Q_2(s_2, noop, a_2'') = \begin{cases} \frac{R_2(s_2, noop, a_2'')}{1-\gamma} & \text{if } T(s_2, noop, a_2'') = s_2 \\ \frac{R_2(s_2, noop, a_2'')}{1-\gamma^2} + \frac{\gamma R_2(s_1, a_1, noop)}{1-\gamma^2} & \text{otherwise} \end{cases} \tag{10}$$

Without loss of generality, assume that $Q_2(s_2, noop, a_2) \geq Q_2(s_2, noop, a_2')$. Then, the first player choosing a_1 and the second player choosing a_2 is a Nash equilibrium. ■

Therefore, without loss of generality, we can assume that: $S = \{1, 2\}$, $A_{1,1} = \{send, keep\}$, $A_{1,2} = \{noop\}$, $A_{2,1} = \{noop\}$, and $A_{2,2} = \{send, keep\}$, and the transition function is:

$$T(1, keep, noop) = 1 \tag{11}$$
$$T(1, send, noop) = 2 \tag{12}$$
$$T(2, noop, keep) = 2 \tag{13}$$
$$T(2, noop, send) = 1 \tag{14}$$

This is the structure of the game in Figure 1. We will call such games **small games**.

Since we will be considering turn-taking games, we will abbreviate R such that:

$$R_i(1, keep) = R_i(1, keep, noop) \quad (15)$$
$$R_i(1, send) = R_i(1, send, noop) \quad (16)$$
$$R_i(2, keep) = R_i(2, noop, keep) \quad (17)$$
$$R_i(2, send) = R_i(2, noop, send) \quad (18)$$

We will define a function $E_\Gamma : \{1, 2\} \times \{send, keep\}^2$, where $E_\Gamma(i, a, a')$ represents the expected utility of agent i in state i (i.e., 1 in 1 or 2 in 2) when a is played in state 1 and a' is played in state 2. Explicitly:

$$E_\Gamma(1, keep, keep) = \frac{1}{1-\gamma} R_1(1, keep) \quad (19)$$
$$E_\Gamma(1, send, keep) = R_1(1, send) + \frac{\gamma}{1-\gamma} R_1(2, keep) \quad (20)$$
$$E_\Gamma(1, keep, send) = \frac{1}{1-\gamma} R_1(1, keep) \quad (21)$$
$$E_\Gamma(1, send, send) = \frac{1}{1-\gamma^2} R_1(1, send) + \frac{\gamma}{1-\gamma^2} R_1(2, send) \quad (22)$$
$$E_\Gamma(2, keep, keep) = \frac{1}{1-\gamma} R_2(2, keep) \quad (23)$$
$$E_\Gamma(2, keep, send) = R_2(2, send) + \frac{\gamma}{1-\gamma} R_2(1, keep) \quad (24)$$
$$E_\Gamma(2, send, keep) = \frac{1}{1-\gamma} R_2(2, keep) \quad (25)$$
$$E_\Gamma(2, send, send) = \frac{1}{1-\gamma^2} R_2(2, send) + \frac{\gamma}{1-\gamma^2} R_2(1, send) \quad (26)$$

Lemma 4 establishes the game in Figure 1 is in fact a NoSDE game.

Lemma 4 *A small game is a NoSDE game if:*

$$E_\Gamma(2, send, send) > E_\Gamma(2, send, keep) \quad (27)$$
$$E_\Gamma(2, keep, keep) > E_\Gamma(2, keep, send) \quad (28)$$
$$E_\Gamma(1, send, keep) > E_\Gamma(1, keep, keep) \quad (29)$$
$$E_\Gamma(1, keep, send) > E_\Gamma(1, send, send) \quad (30)$$

Proof: Each of the four inequalities eliminates one of the four deterministic equilibria.

1. **1 sends, 2 keeps**: Equation 27 states that, if 1 always sends, then the expected utility for 2 always sending is higher than that of 2 always keeping.

2. **1 keeps, 2 sends**: Equation 28 states that, if 1 always keeps, then the expected utility for 2 always keeping is higher than that of 2 always sending.

3. **1 keeps, 2 keeps**: Equation 29 states that, if 2 always keeps, then the expected utility for 1 always sending is higher than that of 1 always keeping.

4. **1 sends, 2 sends**: Equation 30 states that, if 2 always sends, then the expected utility for 1 always keeping is higher than that of 1 always sending.

■

We will call the set of NoSDE games specified in Lemma 4 **type I NoSDE games**.

Lemma 5 characterizes another set of NoSDE games, which we will refer to as **type II NoSDE games**, where the roles of 1 and 2 are reversed (when 2 sends, 1 wants to send, et cetera). The set of type II NoSDE games is isomorphic to the set of type I NoSDE games.

Lemma 5 *A small game is a NoSDE game if:*

$$\begin{aligned}E_\Gamma(1, send, send) &> E_\Gamma(1, keep, send) &(31)\\ E_\Gamma(1, keep, keep) &> E_\Gamma(1, send, keep) &(32)\\ E_\Gamma(2, keep, send) &> E_\Gamma(2, keep, keep) &(33)\\ E_\Gamma(2, send, keep) &> E_\Gamma(2, send, send) &(34)\end{aligned}$$

In Lemma 7 we show that these two sets completely characterize the set of all NoSDE games.

Lemma 6 *For a small game Γ, if any of the following Equations hold:*

$$\begin{aligned}E_\Gamma(2, send, send) &= E_\Gamma(2, send, keep) &(35)\\ E_\Gamma(2, keep, keep) &= E_\Gamma(2, keep, send) &(36)\\ E_\Gamma(1, send, keep) &= E_\Gamma(1, keep, keep) &(37)\\ E_\Gamma(1, keep, send) &= E_\Gamma(1, send, send) &(38)\end{aligned}$$

then Γ is not a NoSDE game.

Proof: We prove this for Equation 35, and the rest are symmetric. Observe that if 2 is indifferent between his actions when 1 plays send, then one of the following is true:

1. **1 keeping is a dominant action if 2 keeps or sends.** Thus, 1 keeping and 2 playing a deterministic best response is a Nash equilibrium, and the game is not a NoSDE game.

2. **1 sending is a best response to 2 keeping.** 1 sending and 2 keeping is a Nash equilibrium, and the game is not a NoSDE game.

3. **1 sending is a best response to 2 sending.** 2 sending and 1 sending is a Nash equilibrium, and the game is not a NoSDE game.

Similar arguments apply to the remaining three equations. ∎

Lemma 7 *A NoSDE game Γ is either a type I NoSDE game or a type II NoSDE game.*

Proof: Consider the relationship between $E_\Gamma(2, send, send)$ and $E_\Gamma(2, send, keep)$.

1. If $E_\Gamma(2, send, send) = E_\Gamma(2, send, keep)$, then by Lemma 6, Γ is not a NoSDE game.

2. If $E_\Gamma(2, send, send) > E_\Gamma(2, send, keep)$, then $E_\Gamma(1, keep, send) > E_\Gamma(1, send, send)$, or 1 sending and 2 sending is a Nash equilibrium. Thus, $E_\Gamma(2, keep, keep) > E_\Gamma(2, keep, send)$, or 1 keeping and 2 sending is a Nash equilibrium. Finally, $E_\Gamma(1, send, keep) > E_\Gamma(1, keep, keep)$, or 1 keeping and 2 keeping is a Nash equilibrium. Thus, if Γ is a NoSDE game, it is a type I NoSDE game.

3. If $E_\Gamma(2, send, keep) > E_\Gamma(2, send, send)$, then $E_\Gamma(1, keep, keep) > E_\Gamma(1, send, keep)$, or 1 sending and 2 keeping is a Nash equilibrium. Thus,

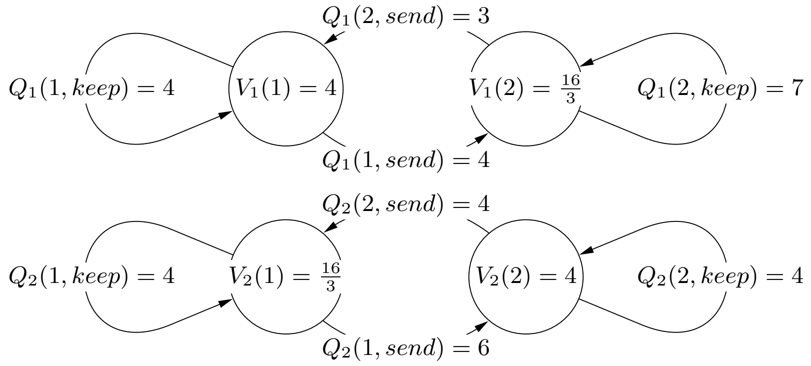

Figure 3: Q and V values for the unique equilibrium in the game in Figure 1.

$E_\Gamma(2, keep, send) > E_\Gamma(2, keep, keep)$, or 1 keeping and 2 keeping is a Nash equilibrium. Finally, $E_\Gamma(1, send, send) > E_\Gamma(1, keep, send)$, or 1 keeping and 2 sending is a Nash equilibrium. Therefore, it is a type II NoSDE game.

∎

Lemma 8 *If 1 randomizes in a small game, then:*

$$V_1(1) = Q_1(1, keep) = Q_1(1, send) = \frac{R_1(1, keep)}{1-\gamma} \tag{39}$$

$$V_1(2) = \frac{1}{\gamma(1-\gamma)} R_1(1, keep) - \frac{1}{\gamma} R_1(1, send) \tag{40}$$

$$Q_1(2, send) = R_1(2, send) + \frac{\gamma}{1-\gamma} R_1(1, keep) \tag{41}$$

$$Q_1(2, keep) = R_1(2, keep) + \frac{R_1(1, keep)}{1-\gamma} - R_1(1, send) \tag{42}$$

Proof: If 1 randomizes, she is indifferent between her actions:

$$V_1(1) = Q_1(1, keep) = Q_1(1, send) \tag{43}$$

$V_1(1)$ can be computed without knowing the precise behavior of 1 and 2, because:

$$Q_1(1, keep) = R_1(1, keep) + \gamma V_1(1) \tag{44}$$

$$V_1(1) = R_1(1, keep) + \gamma V_1(1) \tag{45}$$

$$V_1(1) = \frac{R_1(1, keep)}{1-\gamma} \tag{46}$$

We can solve for $Q_1(2, send)$ now.

$$Q_1(2, send) = R_1(2, send) + \gamma V_1(1) \tag{47}$$

$$Q_1(2, send) = R_1(2, send) + \frac{\gamma}{1-\gamma} R_1(1, keep) \tag{48}$$

We can use the definition of $Q_1(1, send)$ to derive a value for $V_1(2)$:

$$Q_1(1, send) = R_1(1, send) + \gamma V_1(2) \tag{49}$$

$$\frac{R_1(1, keep)}{1-\gamma} = R_1(1, send) + \gamma V_1(2) \tag{50}$$

$$V_1(2) = \frac{1}{\gamma(1-\gamma)} R_1(1, keep) - \frac{1}{\gamma} R_1(1, send) \tag{51}$$

Being aware of $V_1(2)$, we can write an expression for $Q_1(2, keep)$:

$$\begin{aligned} Q_1(2, keep) &= R_1(2, keep) + \gamma V_1(2) && (52) \\ &= R_1(2, keep) + \frac{R_1(1, keep)}{1-\gamma} - R_1(1, send) && (53) \end{aligned}$$

■

Lemma 9 *If 2 randomizes in a small game, then:*

$$\begin{aligned} V_2(2) &= Q_2(2, keep) = Q_2(2, send) = \frac{R_2(2, keep)}{1-\gamma} && (54) \\ V_2(1) &= \frac{1}{\gamma(1-\gamma)} R_2(2, keep) - \frac{1}{\gamma} R_2(2, send) && (55) \\ Q_2(1, send) &= R_2(1, send) + \frac{\gamma}{1-\gamma} R_2(2, keep) && (56) \\ Q_2(1, keep) &= R_2(1, keep) + \frac{R_2(2, keep)}{1-\gamma} - R_2(2, send) && (57) \end{aligned}$$

The proof is symmetric to the proof of Lemma 8.

Lemma 10 *If 1 randomizes in a Nash equilibrium π in a NoSDE game Γ, then 2 will randomize in π in Γ, because $Q_1(2, send) \neq V_1(2)$ and $Q_1(2, keep) \neq V_1(2)$.*

Proof: We will prove that if $Q_1(2, keep) = V_1(2)$, then Γ is not a NoSDE game. From Equation 2:

$$\begin{aligned} Q_1(2, keep) &= R_1(2, keep) + \gamma V_1(2) && (58) \\ V_1(2) &= R_1(2, keep) + \gamma V_1(2) && (59) \\ V_1(2) &= \frac{R_1(2, keep)}{1-\gamma} && (60) \end{aligned}$$

Now, from Equation 40:

$$\begin{aligned} V_1(2) &= \frac{1}{\gamma(1-\gamma)} R_1(1, keep) - \frac{1}{\gamma} R_1(1, send) && (61) \\ \frac{1}{1-\gamma} R_1(2, keep) &= \frac{1}{\gamma(1-\gamma)} R_1(1, keep) - \frac{1}{\gamma} R_1(1, send) && (62) \\ R_1(1, send) + \frac{\gamma}{1-\gamma} R_1(2, keep) &= \frac{R_1(1, keep)}{1-\gamma} && (63) \\ E_\Gamma(1, send, keep) &= E_\Gamma(1, keep, keep) && (64) \end{aligned}$$

Thus, by Equation 37 in Lemma 6, Γ would not be a NoSDE game.

Now, we will prove that if $Q_1(2, send) = V_1(2)$, then Γ is not a NoSDE game. Let us suppose $Q_1(2, send) = V_1(2)$. By Lemma 8:

$$Q_1(2, send) = V_1(2) \quad (65)$$

$$R_1(2, send) + \frac{\gamma}{1-\gamma} R_1(1, keep) = \frac{1}{\gamma(1-\gamma)} R_1(1, keep) - \frac{1}{\gamma} R_1(1, send) \quad (66)$$

$$\frac{1}{1-\gamma^2} R_1(1, send) + \frac{\gamma}{1-\gamma^2} R_1(2, send) = \frac{R_1(1, keep)}{1-\gamma} \quad (67)$$

$$E_\Gamma(1, send, send) = E_\Gamma(1, keep, send) \quad (68)$$

Thus, by Equation 38 in Lemma 6, Γ would not be a NoSDE game.

Finally, we show that when 1 randomizes, 2 randomizes. Since $Q_1(2, keep) \neq V_1(2)$, the strategy of B cannot be to always keep. Similarly, since $Q_1(2, send) \neq V_1(2)$, the strategy of B cannot be to always send. Thus, 2 must randomize in π. ∎

Lemma 11 *If 2 randomizes in a Nash equilibrium π in a NoSDE game Γ, then 1 will randomize in π in Γ, because $Q_2(1, send) \neq V_2(1)$ and $Q_2(1, keep) \neq V_2(1)$.*

The proof is symmetric to the one above.

Lemma 12 *There is a unique Nash equilibrium π in a NoSDE game, and:*

1. *in this Nash equilibrium both agents randomize,*
2. *Q_1 and V_1 are as specified in Lemma 8,*
3. *Q_2 and V_2 are as specified in Lemma 9, and*
4.
$$\pi(1)(send) = \frac{V_2(1) - Q_2(1, keep)}{Q_2(1, send) - Q_2(1, keep)} \tag{69}$$
$$\pi(1)(keep) = \frac{V_2(1) - Q_2(1, send)}{Q_2(1, keep) - Q_2(1, send)} \tag{70}$$
$$\pi(2)(send) = \frac{V_1(2) - Q_1(2, keep)}{Q_1(2, send) - Q_1(2, keep)} \tag{71}$$
$$\pi(2)(keep) = \frac{V_1(2) - Q_1(2, send)}{Q_1(2, keep) - Q_1(2, send)} \tag{72}$$

Proof: Observe that there exists a Nash equilibrium in any stochastic game. Consider an arbitrary equilibrium π. Since π is a Nash equilibrium in a NoSDE game, then at least one agent randomizes. By Lemmas 10 and 11, the other agent must randomize. Therefore, Q_1 and V_1 from Lemma 8 and Q_2 and V_2 from Lemma 9 are the Q and V values for π.

Thus, since all Nash equilibria for some NoSDE game Γ have the same Q and V values, they must all satisfy the equations:

$$V_1(2) = \pi(2)(send)Q_1(2, send) + \pi(2)(keep)Q_1(2, keep) \tag{73}$$
$$V_2(1) = \pi(1)(send)Q_2(1, send) + \pi(1)(keep)Q_2(1, keep) \tag{74}$$
$$1 = \pi(2)(send) + \pi(2)(keep) \tag{75}$$
$$1 = \pi(1)(send) + \pi(1)(keep) \tag{76}$$

Now, if these equations have more than one solution, then one of the following holds:

$$V_1(2) = Q_1(2, send) = Q_1(2, keep) \tag{77}$$
$$V_2(1) = Q_2(1, send) = Q_2(1, keep) \tag{78}$$

. However, if it is a NoSDE game, then by Lemmas 10 and 11, neither Equation 77 nor Equation 78 holds. Thus, the solution is unique. ∎

Lemma 1 follows directly from Lemma 12.

B Proof of Theorem 1

Theorem 5 *A small game can be represented by a vector in $\mathbf{R}^{n \times \mathcal{A}}$, where each coordinate represents a reward for some player when some action is taken in some state. In this representation, the set of all NoSDE games for a fixed γ is an open set.*

Proof: Observe that the set of type I NoSDE games are represented by a system of strict linear inequalities, and therefore it is an open set. Similarly, the set of type II NoSDE games are represented by a system of strict linear inequalities, and therefore it is an open set. The union of these two sets is also open.

■

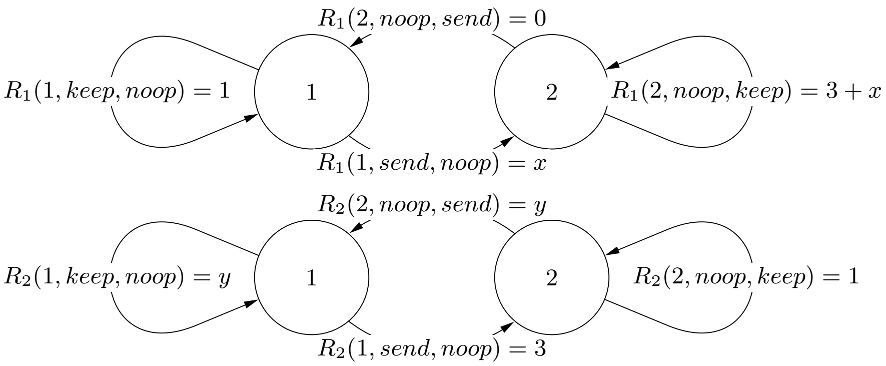

Figure 4: A modification of the rewards in the WOD game in Figure 1 which yields different equilibria but the same Q-values for x in the open interval $\left(-\frac{5}{4}, \frac{7}{4}\right)$ (where $Q_1(2, send) < V_1(2) < Q_1(2, keep)$, see Figure 5) and y in the open interval $\left(-\frac{1}{2}, 1\right)$.

Lemma 13 *If* $\Gamma = (S, A, T, R)$ *is a NoSDE game, and* $\Gamma' = (S, A, T, R')$ *is a WOD game, and there exists an* x *such that:*

$$R'_i(s, a) = \begin{cases} R_i(s, a) + x \text{ if } i = 1, s = 1, a = send \\ R_i(s, a) + x \text{ if } i = 1, s = 2, a = keep \\ R_i(s, a) \qquad \text{otherwise.} \end{cases} \tag{79}$$

then the equilibrium of Γ *and the equilibrium of* Γ' *have identical* Q*-values, and if* x *is nonzero, they have different* V*-values and different equilibria.*

Proof: Observe from Lemma 8 that Q_1 and V_1 are not dependent upon R_2, and symmetrically Q_2 and V_2 are not dependent upon R_1. Thus, we can assume that Q_2 and V_2 for Γ and Γ' are equal. Observe from Equations 39 and 41 that $V_1(1)$, $Q_1(1, keep)$, $Q_1(1, send)$, and $Q_1(2, send)$ do not depend upon $R_1(2, keep)$ or $R_1(1, send)$. Moreover, by Equation 42, the effect of increasing $R_1(2, keep)$ and $R_1(1, send)$ upon $Q_1(2, keep)$ exactly cancel out. Finally, by Equation 40, increasing $R_1(1, send)$ by x decreases $V_1(2)$ by $\frac{x}{\gamma}$. ■

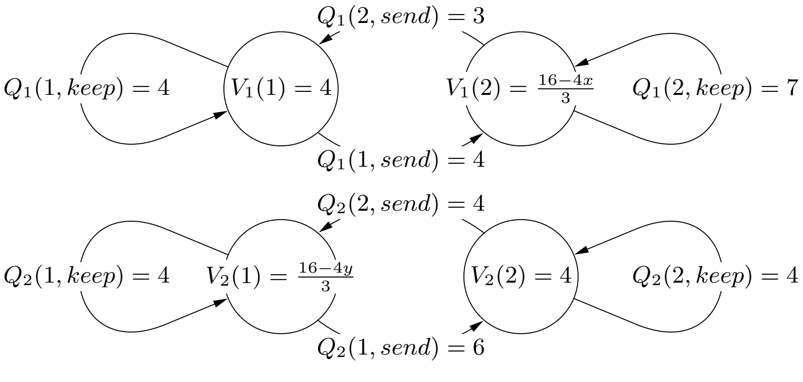

Figure 5: Q and V values for the unique equilibrium in the game in Figure 4.

By the symmetry between 1 and 2, we can also modify R_2 in a symmetric fashion.

Proof (of Theorem 1): Observe that by Theorem 5, for any NoSDE game $\Gamma = (S, A, T, R)$, there exists an $\epsilon > 0$ such that for all games $\Gamma' = (S, A, T, R')$ where $|R_i(s,a) - R'_i(s,a)| < \epsilon$, Γ' is also a NoSDE game. By Lemma 13, some of these games have the same equilibrium Q-values, but different equilibrium V-values and different equilibria. ∎

C Proof of Theorem 2

Define $\mathcal{V}$ to be the space of all state value functions and $\mathcal{Q}$ to be the space of all state-action value functions.

For all $\pi : S \to A$, define $V^{\pi}_{\mathcal{Q}} : \mathcal{Q} \to \mathcal{V}$, $\mathcal{Q}_{\mathcal{V}} : \mathcal{V} \to \mathcal{Q}$, and $\mathcal{V}^{\pi}_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}$ to be:

$$(\mathcal{V}^{\pi}_{\mathcal{Q}}(Q))_i(s) = Q_i(s, \pi(s)) \tag{80}$$

$$(\mathcal{Q}_{\mathcal{V}}(V))_i(s,a) = R_i(s,a) + \gamma \sum_{s' \in S} T(s,a)(s')V(s') \tag{81}$$

$$\mathcal{V}^{\pi}_{\mathcal{V}} = V^{\pi}_{\mathcal{Q}} \circ \mathcal{Q}_{\mathcal{V}} \tag{82}$$

Define the value differences to be:

$$\delta V^t_1 = V^t_1(1) - V^t_1(2) \tag{83}$$

$$\delta V^t_2 = V^t_2(1) - V^t_2(2) \tag{84}$$

We can describe the util-CE update rule in terms of only these differences.

Lemma 14 *For some policy π, if $V_t = \mathcal{V}^{\pi}_{\mathcal{V}}(V_{t-1})$, then for any $P \in \{1, 2\}$:*

1. *If $\pi(1) = \pi(2) = send$:*
$$\delta V^t_P = R_P(1, send) - R_P(2, send) - \gamma \delta V^{t-1}_P \tag{85}$$

2. *If $\pi(1) = \pi(2) = keep$:*
$$\delta V^t_P = R_P(1, keep) - R_P(2, keep) + \gamma \delta V^{t-1}_P \tag{86}$$

3. *If $\pi(1) = keep$ and $\pi(2) = send$:*
$$\delta V^t_P = R_P(1, keep) - R_P(2, send) \tag{87}$$

4. *If $\pi(1) = send$ and $\pi(2) = keep$:*
$$\delta V^t_P = R_P(1, send) - R_P(2, keep) \tag{88}$$

The proof follows directly from the update rule. Now, let us consider the crossover points between where the agents select different policies.

$$\delta V^*_1 = \frac{1}{\gamma}\left(R_1(1, send) - R_1(1, keep)\right) \tag{89}$$

$$\delta V^*_2 = \frac{1}{\gamma}\left(R_2(2, keep) - R_2(2, send)\right) \tag{90}$$

In the game in Figure 1, $\delta V^*_1 = -\frac{4}{3}$, and $\delta V^*_2 = \frac{4}{3}$.

Unless one of the value differences is on its crossover point, all equilibrium selection rules are the same. Formally:

Lemma 15 *For any equilibrium update rule:*

1. *If* $\delta V_1^{t-1} > \delta V_1^*$, *then* $\pi^t(1) = keep$.
2. *If* $\delta V_1^{t-1} < \delta V_1^*$, *then* $\pi^t(1) = send$.
3. *If* $\delta V_2^{t-1} > \delta V_2^*$, *then* $\pi^t(2) = send$.
4. *If* $\delta V_2^{t-1} < \delta V_2^*$, *then* $\pi^t(2) = keep$.

This can be proven by simply computing Q_t. For the case of util-CE, we can also make the following assertion. First of all, we define secondary crossover points, for the agent that is not in control.

$$D_2 = \frac{1}{\gamma}\left(R_2(1, send) - R_2(1, keep)\right) \tag{91}$$

$$D_1 = \frac{1}{\gamma}\left(R_1(2, keep) - R_1(2, send)\right) \tag{92}$$

Lemma 16 *For the util-CE update rule:*

1. *If* $\delta V_1^{t-1} = \delta V_1^*$, *and* $\delta V_2^{t-1} < D_2$, *then* $\pi^t(1) = send$.
2. *If* $\delta V_1^{t-1} = \delta V_1^*$, *and* $\delta V_2^{t-1} > D_2$, *then* $\pi^t(1) = keep$.
3. *If* $\delta V_2^{t-1} = \delta V_2^*$, *and* $\delta V_1^{t-1} > D_1$, *then* $\pi^t(2) = send$.
4. *If* $\delta V_2^{t-1} = \delta V_2^*$, *and* $\delta V_1^{t-1} < D_1$, *then* $\pi^t(2) = keep$.

Proof: This follows directly from the fact that if the player with more than one action in a state is indifferent between her actions, util-CE maximizes the utility of the player that has only one action in that state. ∎

Finally, if both agents are indifferent for the action for a state, it does not matter what happens.

Lemma 17 *For any equilibrium update rule:*

1. *If* $\delta V_1^{t-1} = \delta V_1^*$, *and* $\delta V_2^{t-1} = D_2$, *then for any policy* π, *for any player* $P \in \{1, 2\}$, $\mathcal{V}_{\mathcal{V}}^{\pi}(V_{t-1})_P(1) = V_P^t(1)$.
2. *If* $\delta V_2^{t-1} = \delta V_2^*$, *and* $\delta V_1^{t-1} = D_1$, *then for any policy* π, *for any player* $P \in \{1, 2\}$, $\mathcal{V}_{\mathcal{V}}^{\pi}(V_{t-1})_P(2) = V_P^t(2)$.

Corollary 1 *For the util-CE update rule:*

1. *If the first player randomizes in* π^t, *then* $\delta V_1^{t-1} = \delta V_1^*$, $\delta V_2^{t-1} = D_2$, *and for any policy* π, *for any player* $P \in \{1, 2\}$, $\mathcal{V}_{\mathcal{V}}^{\pi}(V_{t-1})_P(1) = V_P^t(1)$.
2. *If the second player randomizes in* π^t, *then* $\delta V_2^{t-1} = \delta V_2^*$, $\delta V_1^{t-1} = D_1$, *and for any policy* π, *for any player* $P \in \{1, 2\}$, $\mathcal{V}_{\mathcal{V}}^{\pi}(V_{t-1})_P(2) = V_P^t(2)$.

Proof: If both players are indifferent to the choice of policy for a state, then the choice of policy has no effect on the next value. ∎

Lemma 18 *For any equilibrium update rule, if there exists a* k^*, T *such that for all* $t > T$, *there exists a* π *such that* $\mathcal{V}_{\mathcal{V}}^{\pi}(V_t) = V_{t+1}$ *and* $\mathcal{V}_{\mathcal{V}}^{\pi}(V_{t+k^*}) = V_{t+k^*+1}$ *then there exists a*

T' such that for all $T'' > T'$, GetCycle(Γ,V^0,f,T'',$\lceil T''/2 \rceil$) returns a cyclic correlated equilibrium.

Proof: For all $i \in \{1, \ldots, k^*\}$, define $\rho^i = \rho^{T+i+1}$. Therefore, $\mathcal{V}_{\mathcal{V}}^{\rho^i}(V_{T+i}) = V_{T+i+1}$. Then we know that, for any non-negative integer j, $\mathcal{V}_{\mathcal{V}}^{\rho^i}(V_{T+i+jk^*}) = V_{T+i+jk^*+1}$. Define $f_i : \mathcal{V} \to \mathcal{V}$ such that:

$$f_i = \mathcal{V}_{\mathcal{V}}^{\rho^{i-1}} \circ \mathcal{V}_{\mathcal{V}}^{\rho^{i-2}} \circ \ldots \circ \mathcal{V}_{\mathcal{V}}^{\rho^1} \circ \mathcal{V}_{\mathcal{V}}^{\rho^{k^*}} \circ \mathcal{V}_{\mathcal{V}}^{\rho^{k^*-1}} \circ \ldots \circ \mathcal{V}_{\mathcal{V}}^{\rho^i} \tag{93}$$

Therefore, $f_i(V_{T+i+jk^*}) = V_{T+i+(j+1)k^*}$ for all non-negative j. Also, observe that f_i is a contraction mapping, because for any ρ, $\mathcal{V}_{\mathcal{V}}^{\rho}$ is a contraction mapping. Define $V^{*,i}$ to be the fixed point of f_i. Observe that:

$$\lim_{j \to \infty} V_{T+i+jk^*} = V^{*,i} \tag{94}$$

Moreover, by definition, $\mathcal{V}_{\mathcal{V}}^{\rho^i}(V^{*,i}) = V^{*,inc_{k^*}(i)}$. Thus, $V^{k^*}, \ldots, V^1$ are the value functions of a cyclic correlated equilibrium. This implies that (ρ^{k^*},...,ρ^1) is a cyclic correlated equilibrium.

Now, we prove that the GetCycle algorithm actually extracts an equilibrium from this sequence. Define $i \cong i'$ if and only if $V^{*,i} = V^{*,i'}$. Define $\epsilon > 0$ such that for all i, i' where $i \ncong i'$, $d(V^{*,i}, V^{*,i'}) > \epsilon$. Define J such that for all $j > J$, for all $i \in \{1 \ldots k^*\}$, $d(V^{T+i+jk^*}, V^{*,i}) < \frac{\epsilon}{4}$. For $i, i' \in \{1 \ldots k^*\}$ where $i \ncong i'$, for all $j, j' > J$, this implies that $d(V^{T+i+jk^*}, V^{*,i'}) > \frac{3\epsilon}{4}$ and that $d(V^{T+i+jk^*}, V^{T+i'+j'k^*}) > \frac{\epsilon}{2}$. It also implies that $d(V^{T+i+jk^*}, V^{T+i+j'k^*}) < \frac{\epsilon}{2}$.

Define $T' = 2T + 2Jk^* + 4$. Observe that $T' - \lceil T'/2 \rceil \geq T + Jk^* + 1$. Fix some $T'' > T'$ where $T'' = T + j''k^* + i''$. Thus, when k is selected in Step 3 of GetCycle, it will only consider value functions larger than $T + Jk^* + 1$. Thus, if we define $i^* \in \{1 \ldots k\}$ and j^* as an integer such that $T + j^*k^* + i^* = T'' - k$. By the previous paragraph, we know that $i^* \cong i''$, so $V^{*,i^*} = V^{*,i''}$. Therefore, the sequence of policies $\pi^{T''}, \ldots, \pi^{T+j^*k^*+i^*+1}$ is a cyclic equilibrium. ■

Lemma 19 *If one uses the util-CE update rule, and there exists t, t' such that $\delta V_1^t = \delta V_1^{t'}$ and $\delta V_2^t = \delta V_2^{t'}$, then there exists a π such that $\mathcal{V}_{\mathcal{V}}^{\pi}(V_t) = V_{t+1}$, $\mathcal{V}_{\mathcal{V}}^{\pi}(V^{t'}) = V^{t'+1}$, $\delta V_1^{t+1} = \delta V_1^{t'+1}$, and $\delta V_2^{t+1} = \delta V_2^{t'+1}$.*

Proof: This follows from Lemma 15, Lemma 16, and Lemma 17. ■

Lemma 20 *If one uses the util-CE update rule, and there exists t, t' such that $\delta V_1^t = \delta V_1^{t'}$ and $\delta V_2^t = \delta V_2^{t'}$, then for some T, for all $T' > T$, GetCycle(Γ,V^0,f,T',$\lceil T'/2 \rceil$) returns a cyclic equilibrium.*

Proof: This follows from Lemma 19 and Lemma 18. ■

Lemma 21 *If one uses the utilCE update rule, and for $t \neq t'$, there exists a π such that $\mathcal{V}_{\mathcal{V}}^{\pi}(V^t) = V^{t+1}$ and $\mathcal{V}^{\pi}(V^{t'}) = V^{t'+1}$, and either:*

1. *$\pi(1) = send$ and $\pi(2) = keep$, or*
2. *$\pi(1) = keep$ and $\pi(2) = send$,*

then there exists a T such that for all $T' > T$ GetCycle(Γ,V^0,f,T',$\lceil T'/2 \rceil$) returns a cyclic equilibrium.

Proof: This follows from Lemma 14 and Lemma 20. ∎

Lemma 22 *For all t, there must exist a $t' > t$ such that either $\pi^{t'}(1) \neq send$ or $\pi^{t'}(2) \neq send$.*

Proof: If this were not the case, then by Lemma 18, π where $\pi(1) = \pi(2) = send$ would be a stationary equilibrium. ∎

Lemma 23 *For all t, there must exist a $t' > t$ such that either $\pi^{t'}(1) \neq keep$ or $\pi^{t'}(2) \neq keep$.*

Proof: Otherwise, π where $\pi(1) = \pi(2) = keep$ is a stationary equilibrium. ∎

Thus, we have established that, if a deterministic policy where the players play different actions is selected twice, then the algorithm will return a cyclic policy. However, it is impossible for both players to play just keep forever or for both players to play just send forever. What remains to be seen if the two agents can, in lock-step, move from both sending to both keeping simultaneously, and then back again in an aperiodic fashion. The next lemma shows that this cannot happen.

Lemma 24 *If $\pi^t(1) = \pi^t(2) = keep$, then $\pi^{t+1}(1) \neq send$ or $\pi^{t+1}(2) \neq send$.*

Proof: We will prove that if the NoSDE game is a type I NoSDE Game, then $\pi^{t+1}(2) = keep$. If the game is a type II NoSDE Game, then $\pi^{t+1}(1) = keep$, for analogous reasons.

By Lemma 15, we know that $\delta V_2^{t-1} \leq \delta V_2^*$. Therefore:

$$\begin{aligned}
\delta V_2^t &= (R_2(1, keep) - R_2(2, keep)) + \gamma \delta V_2^{t-1} && (95)\\
&\leq (R_2(1, keep) - R_2(2, keep)) + \gamma \delta V_2^* && (96)\\
&\leq (R_2(1, keep) - R_2(2, keep)) \\
&\quad + (R_2(2, keep) - R_2(2, send)) && (97)\\
&\leq (R_2(1, keep) - R_2(2, send)) && (98)
\end{aligned}$$

Because this is a type I NoSDE Game:

$$\frac{R_2(2, keep)}{1-\gamma} > \frac{\gamma}{1-\gamma} R_2(1, keep) + R_2(2, send) \qquad (99)$$

$$\frac{1}{\gamma}(R_2(2, keep) - R_2(2, send)) > (R_2(1, keep) - R_2(2, send)) \qquad (100)$$

Therefore:

$$\begin{aligned}
\delta V_2^t &< \frac{1}{\gamma}(R_2(2, keep) - R_2(2, send)) && (101)\\
\delta V_2^t &< \delta V_2^* && (102)
\end{aligned}$$

By Lemma 15, $\pi^{t+1}(2) = keep$. ∎

Lemma 25 *If, for the update rule utilCE, for some t, if neither $\pi^t(1) = \pi^t(2) = keep$ or $\pi^t(1) = \pi^t(2) = send$, then there exists a π where $\mathcal{V}_{\mathcal{V}}^{\pi}(V^{t-1}) = V^t$, where either:*

1. *$\pi(1) = send$ and $\pi(2) = keep$, or*
2. *$\pi(1) = keep$ and $\pi(2) = send$.*

Proof: If both players randomize in π^t, then by Corollary 1, we know that for the policy π where $\pi(1) = keep$ and $\pi(2) = send$, $V^{\pi}_{\mathcal{V}}(V^{t-1}) = V^t$.

If the first player plays any action a deterministically, then the other agent either randomizes or plays the action $a' \neq a$. Thus, for the policy π where $\pi(1) = a$ and $\pi(2) = a'$, $V^{\pi}(V^{t-1}) = V^t$. One can use a similar argument for the case where the second player plays an action deterministically. ■

Lemma 26 *If one uses the utilCE update rule, if there does not exist a t such that for all $t' > t$, either $\pi^{t'}(1) = \pi^{t'}(2) = keep$ or $\pi^{t'}(1) = \pi^{t'}(2) = send$, then there exists a T such that for all $T' > T$, GetCycle(Γ,V^0,f,T',$\lceil T'/2 \rceil$) will return a cyclic equilibrium.*

Proof: Take t_1, t_2, t_3 such that for all i, neither $\pi^{t_i}(1) = \pi^{t_i}(2) = send$ nor $\pi^{t_i}(1) = \pi^{t_i}(2) = keep$, and for all $j \neq i$, $t_i \neq t_j$. These three will exist if there does not exist a t as described in the lemma. For all $i \in \{1, 2, 3\}$, by Lemma 25, we know that we can choose a policy ρ^i where $\mathcal{V}^{\rho^i}(V^{t_i - 1}) = V^{t_i}$ and either $\rho^i(1) = keep$ and $\rho^i(2) = send$, or $\rho^i(1) = send$ and $\rho^i(2) = keep$. Thus, since there are only two possible values of ρ^i and there are three ρ^i, there have to exist i, j where $i \neq j$ and $\rho^i = \rho^j$. By Lemma 21, GetCycle will return an equilibrium for a large enough value of T. ■

Lemma 27 *The assumption in Lemma 26 is true.*

Proof: Let us assume it is false (there is a T such that for all $t > T$, $\pi^t(1) = \pi^t(2) = keep$ or $\pi^t(1) = \pi^t(2) = send$), and prove a contradiction. By Lemma 22, it cannot be the case that for all $t > T$, $\pi^t(1) = \pi^t(2) = send$. Therefore there must exist a $T' > T$ such that $\pi^{T'}(1) = \pi^{T'}(2) = keep$. By our hypothesis that the assumption in Lemma 26 is false and Lemma 24, for all $t > T'$, $\pi^t(1) = \pi^t(2) = keep$. However, this contradicts Lemma 23. Therefore, by contradiction, the assumption in Lemma 26 is true. ■

Proof (of Theorem 2): This follows directly from Lemma 26 and Lemma 27. ■

D Proof of Theorem 3

Proof (of Theorem 3): By Lemma 27, the assumption in Lemma 26 is true. Therefore, for every t, there exists a $t' > t$ such that neither $\pi^{t'}(1) = \pi^{t'}(2) = keep$ nor $\pi^{t'}(1) = \pi^{t'}(2) = send$. By Lemma 25, there exists a π such that $\mathcal{V}^{\pi}_{\mathcal{V}}(V^{t'-1}) = V^{t'}$, where either $\pi(1) = send$ and $\pi(2) = keep$ or $\pi(1) = keep$ and $\pi(2) = send$. By Lemma 14, either:

1. $\delta V^{t'}_P = R_P(1, keep) - R_P(2, send)$, or
2. $\delta V^{t'}_P = R_P(1, send) - R_P(2, keep)$.

For the equilibrium value function V^{π^*}, $V_1^{\pi^*}(1) - V_1^{\pi^*}(2) = \delta V_1^*$, and $V_2^{\pi^*}(1) - V_2^{\pi^*}(2) = \delta V_2^*$, because in equilibrium both agents are indifferent between their respective choices.

From the fact that the game has no stationary equilibria, we know the following four inequalities hold:

1. $R_1(1, keep) - R_1(2, send) \neq \delta V_1^*$
2. $R_1(1, send) - R_1(2, keep) \neq \delta V_1^*$
3. $R_2(1, keep) - R_2(2, send) \neq \delta V_2^*$
4. $R_2(1, send) - R_2(2, keep) \neq \delta V_2^*$

Define $\epsilon > 0$ to be strictly less than the absolute difference between the sides of any of the four inequalities above. Therefore, the sequence of value functions will never be ad infinitum closer than ϵ (in the L_1 metric) to V^{π^*}. ■

E Proof of Theorem 4

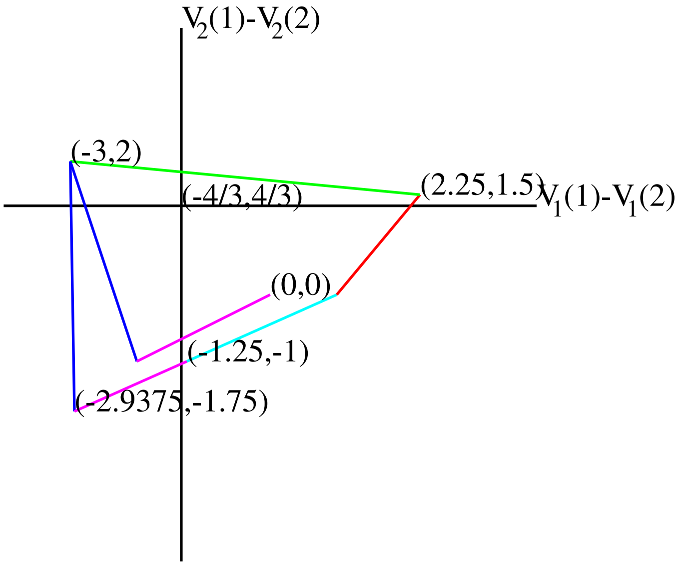

Figure 6: δV_1^t and δV_2^t in the game in Figure 1 when $V^0 = 0$. Observe that the points never touch a boundary δV_1^* or δV_2^*, and that they reach the same point twice.

Proof (of Theorem 4): In Figure 6, we observe how δV_1^t and δV_2^t change when the utilCE update rule is used seven times. However, observe that for all $t \in \{1, \ldots, 7\}$, $\delta V_1^t \neq \delta V_1^* = -4/3$ and $\delta V_2^t \neq \delta V_2^* = 4/3$, by Lemma 15, all equilibrium update rules will behave the same. Moreover, by Lemma 14, these differences will cycle exactly and forever. Thus, by an argument similar to the proof of Lemma 18, the sequence of value functions will converge to a limit cycle, i.e., they will not converge to any particular value. ■

F Answer to Sergiu Hart's Question

One natural question is whether or not the existence of NoSDE games depends on whether the rewards and gamma are rational or not. While Theorem 5 establishes that NoSDE games can have irrational rewards, this section shows that there are NoSDE games for all $\gamma \in (0, 1)$.

Lemma 28 *For any small game Γ with utility R, there exists an N such that for any $x > N$,* *the new small game Γ' with utility R' defined by:*

$$R'_i(s,a) = \begin{cases} R_i(s,a) - x \text{ if } i=1, s=2, a=send \\ R_i(s,a) + x \text{ if } i=1, s=2, a=keep \\ R_i(s,a) + x \text{ if } i=2, s=1, a=send \\ R_i(s,a) - x \text{ if } i=2, s=1, a=keep \\ R_i(s,a) \quad\quad otherwise, \end{cases} \tag{103}$$

then Γ' is a ***type I NoSDE game****.*

Corollary 2 *For all $\gamma \in (0,1)$, there exists a NoSDE game with a discount factor of γ.*

G Stronger Non-Convergence Result

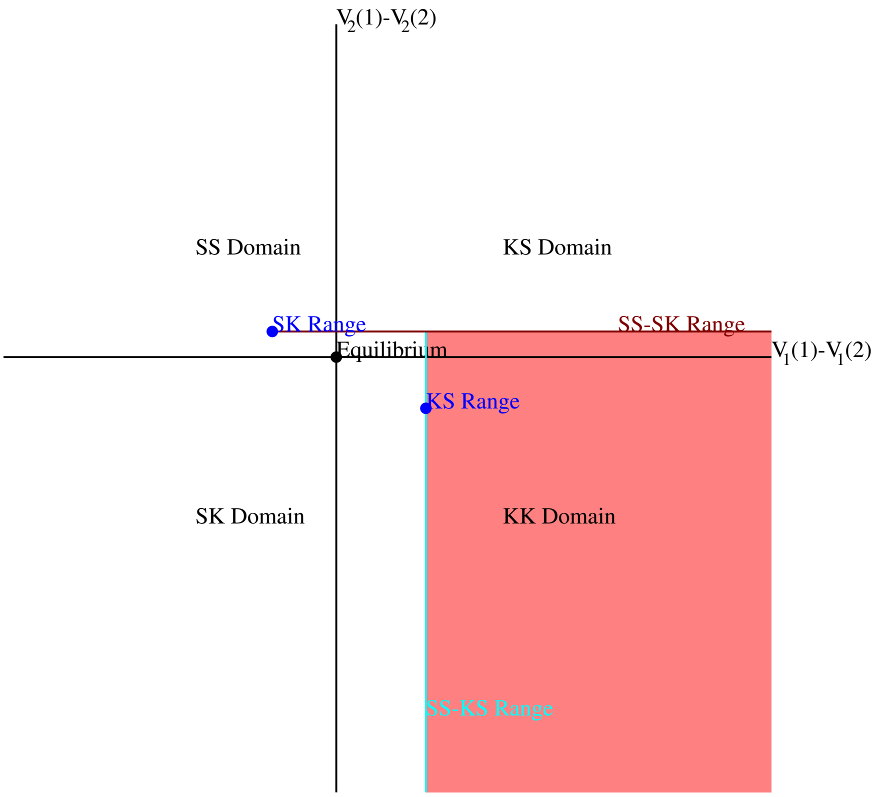

Figure 7: The range of δV_1^t and δV_2^t in the game in Figure 1 when $\delta V_1^{t-1} \leq \delta V_1^*$ and $\delta V_2^{t-1} \geq \delta V_2^*$.

Define:

1. $\delta V_1^{sk} = R_1(1, send) - R_1(2, keep)$
2. $\delta V_2^{sk} = R_2(1, send) - R_2(2, keep)$
3. $\delta V_1^{ks} = R_1(1, keep) - R_1(2, send)$
4. $\delta V_2^{ks} = R_2(1, keep) - R_2(2, send)$

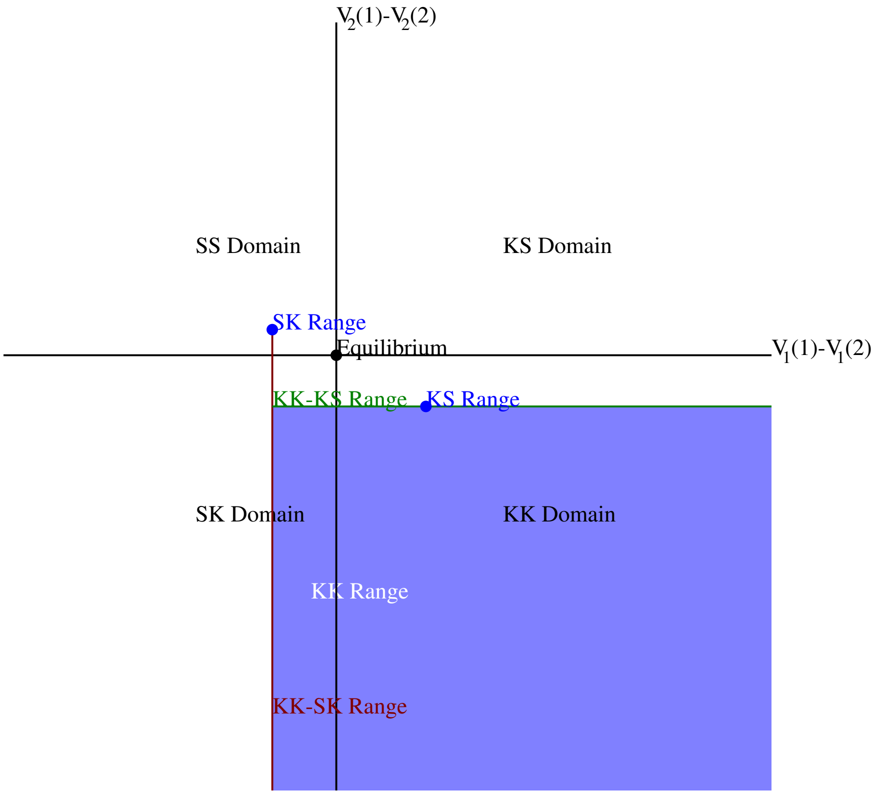

Figure 8: The range of δV_1^t and δV_2^t in the game in Figure 1 when $\delta V_1^{t-1} \geq \delta V_1^*$ and $\delta V_2^{t-1} \leq \delta V_2^*$.

Lemma 29 *For any Type I NoSDE Game,*

1. $\delta V_1^{sk} < \delta V_1^*$
2. $\delta V_2^{sk} > \delta V_2^*$
3. $\delta V_1^{ks} > \delta V_1^*$
4. $\delta V_2^{ks} < \delta V_2^*$

Proof: Part 1 follows from Equation 29, Part 2 follows from Equation 27, Part 3 follows from Equation 30, and Part 4 follows from 28. ∎

Also, arithmetic shows that:[4]

Fact 3 *For any NoSDE game:*

1. $R_1(1, keep) - R_1(2, keep) + \gamma(\delta V_1^*) = \delta V_1^{sk}$.
2. $R_1(1, send) - R_1(2, send) - \gamma(\delta V_1^*) = \delta V_1^{ks}$.
3. $R_2(1, keep) - R_2(2, keep) + \gamma(\delta V_2^*) = \delta V_2^{ks}$.
4. $R_2(1, send) - R_2(2, send) - \gamma(\delta V_2^*) = \delta V_2^{sk}$.

Define $\delta V^{\min} = \min(|\delta V_1^{sk} - \delta V_1^*|, |\delta V_1^{ks} - \delta V_1^*|, |\delta V_2^{sk} - \delta V_2^*|, |\delta V_2^{ks} - \delta V_2^*|)$. By Lemma 29, $\delta V^{\min} > 0$ for Type I NoSDE Games, and one can prove $\delta V^{\min} > 0$ for Type II NoSDE Games in a similar fashion.

Now, we will give formal statements associated with Figures 7 and 8.

Lemma 30 *For any Type I NoSDE Game Γ, for any equilibrium update rule $f \in F_\Gamma$, if $\delta V^t = \delta V^*$, then $d(\delta V^{t+1}, \delta V^*) \geq \delta V^{\min} > 0$.*

Basically, if at least one player is not indifferent given V^t, then the other player's utility is the same regardless of what equilibrium update rule is used, and we can bound this away from the equilibrium utility.

Proof: If any equilibrium update rule is used, then:

1. If $\delta V_1^t > \delta V_1^*$ (first player sends):
   (a) If $\delta V_2^t < \delta V_2^*$ (second player sends): then $\delta V_2^{t+1} > R_2(1, send) - R_2(2, send) - \gamma(\delta V_2^*) = \delta V_2^{sk}$. By Lemma 29, $\delta V_2^{t+1} > \delta V_2^{sk} > \delta V_2^*$.
   (b) $\delta V_2^t > \delta V_2^*$ (Second player keeps): then $\delta V^{t+1} = \delta V^{sk} \neq \delta V^*$.
   (c) Second player indifferent ($\delta V_2^t = \delta V_2^*$): then δV_2^{t+1} is in-between δV_2^{sk} and $R_2(1, send) - R_2(2, send) - \gamma(\delta V_2^*)$. But, Fact 3 shows that these are equal, and by definition, $|\delta V_2^{sk} - \delta V_2^*| \geq \delta V^{\min}$, so that $d(\delta V^{t+1}, \delta V^*) > \delta V^{\min}$.
2. $\delta V_1^t < \delta V_1^*$ (First player keeps):
   (a) $\delta V_2^t < \delta V_2^*$ (Second player sends): then $\delta V^{t+1} = \delta V^{ks}$, and $d(\delta V^{ks}, \delta V^*) > \delta V^{\min}$.
   (b) $\delta V_2^t > \delta V_2^*$ (Second player keeps): then $\delta V_1^{t+1} < R_1(1, keep) - R_1(2, keep) + \delta V_1^* = \delta V_1^{sk}$. By Lemma 29, $\delta V_1^{t+1} < \delta V_1^{sk} < \delta V_1^*$.

[4] This can also be derived from the intuition that if an agent is indifferent between its actions, it must obtain the same utility regardless of which it chooses.

(c) $\delta V_2^t = \delta V_2^*$ (Second player indifferent): Similar to when the first player keeps and the second player is indifferent, $\delta V_2^{t+1} = \delta V_2^{ks}$, and $|\delta V_2^{ks} - \delta V_2^*| \geq \delta V^{\min}$.

3. $\delta V_1^t = \delta V_1^*$ (First player indifferent):

   (a) $V_2^t < V_2^*$ (Second player sends): Similar to when the first player keeps and the second player is indifferent, $\delta V_1^{t+1} = \delta V_1^{ks}$, and $|\delta V_1^{ks} - \delta V_1^*| \geq \delta V^{\min}$.

   (b) $V_2^t > V_2^*$ (Second player keeps): Similar to when the first player keeps and the second player is indifferent, $V_1^{t+1} = V_1^{sk}$, and $|\delta V_1^{sk} - \delta V_1^*| \geq \delta V^{\min}$.

   (c) $(\delta V_2^t = \delta V_2^*)$ (Second player indifferent): By assumption, it is not the case that both agents are indifferent.

■

Similarly, we can prove:

Lemma 31 *For any Type II NoSDE Game Γ, for any equilibrium update rule $f \in F_\Gamma$, if $\delta V^t = \delta V^*$, then $d(\delta V^{t+1}, \delta V^*) \geq \delta V^{\min} > 0$.*

Now, given these results about updates, we can make a statement about the behavior as a whole.

Theorem 6 *Given any NoSDE game Γ with equilibrium π and **any** equilibrium update rule $f \in F_\Gamma$, unless there exists $c_1, c_2 \in \mathbf{R}$ such that $V_i^0 = V_i^\pi + c_i$, then if the value-function sequence $\{V^1, V^2, \ldots\}$ is returned from **ValueIteration**(Γ, V^0, f, ∞), for all t, $d(V^t, V^\pi) > \delta V^{\min} > 0$, and therefore the sequence does not converge to V^π.*

For the utilitarian update rule, an even stronger result can be obtained.

Lemma 32 *For any turn-taking game with two players, given any action value function Q^t, all the policies that agree with Q^t that maximize the total utility at every state return the same V^t, and at least one is deterministic.*

Proof: At each state $s \in S$, only one agent needs to make a decision. Let us say only player one needs to make a decision at s. Then $V_1^t(s) = \max_{a \in A_s} Q_1^t(s,a)$, or the policy and Q^t do not agree. However, it might be the case that several actions $a \in A_s$ maximize $Q_1^t(s,a)$, so we can define the set of actions that maximize $Q_1^t(s,a)$ to be A'_s. Now, $V_2^t(s) = \max_{a \in A'_s} Q_2^t(s,a)$, or the policy is not maximizing the total utility. Therefore, for all states, the V^t obtained is unique. Moreover, there must exist an $a_s \in A'_s$ such that $V_i^t(s) = Q_i^t(s,a)$ for $i \in \{1,2\}$, because A'_s is nonempty, and $V_2^t(s)$ is the maximum over a finite set. The policy π where $\pi(s) = a_s$ for all $s \in S$ agrees with Q^t and maximizes total utility among all policies that agree with Q^t. ■

Lemma 33 *If the utilitarian update rule is used, and $\delta V^t = \delta V^*$, then either $\delta V_1^{t+1} = \delta V_1^{ks}$ or $\delta V_1^{t+1} = \delta V_1^{sk}$, and therefore $d(\delta V^{t+1}, \delta V^*) \geq \delta V^{\min}$.*

Proof: If the utilitarian update rule is used, then considering all deterministic policy updates will suffice. Since all of these map δV^* to either δV_1^{ks} or δV_1^{sk} (see Fact 3), the result holds. ■

Theorem 7 *Given any NoSDE game Γ with equilibrium π and the utilitarian update rule $f \in F_\Gamma$ and **any** $V^0 \in \mathcal{V}_\Gamma$, then if the value-function sequence $\{V^1, V^2, \ldots\}$ is returned from **ValueIteration**(Γ, V^0, f, ∞), for all t, $d(V^t, V^\pi) > \delta V^{\min} > 0$, and therefore the sequence does not converge to V^π.*

Proof: This follows from Lemmas 30, 31, and 33. ∎

H Stronger Convergence Result

Theorem 8 *Given the utilitarian-CE equilibrium-selection rule f, for every NoSDE game Γ, for every $V^0 \in \mathcal{V}_\Gamma$, there exists some finite T such that for all $T' > T$,* ***GetCycle***$(\Gamma, V^0, f, T', \lceil T'/2 \rceil)$ *returns a cyclic correlated equilibrium.*

This follows directly from the proof of Theorem 2.