Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Mathematical Psychology 48 (2004) 334–354

Journal of Mathematical Psychology

www.elsevier.com/locate/jmp

Rationality and bounded information in repeated games, with application to the iterated Prisoner's Dilemma

Matt Jones*, Jun Zhang

Department of Psychology, University of Michigan, MI, USA

Received 6 September 2003; received in revised form 23 April 2004

Abstract

Actions in a repeated game can in principle depend on all previous outcomes. Given this vast policy space, human players may often be forced to use heuristics that base actions on incomplete information, such as the outcomes of only the most recent trials. Here it is proven that such bounded rationality is often fully rational, in that the optimal policy based on some limited information about the game's history will be universally optimal (i.e., within the full policy space), provided that one's opponents are restricted to using this same information. It is then shown how this result allows explicit calculation of subgame-perfect equilibria (SPEs) for any repeated or stochastic game. The technique is applied to the iterated Prisoner's Dilemma for the case of 1-back memory. Two classes of SPEs are derived, which exhibit varying degrees of (individually rational) cooperation as a result of repeated interaction.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Game theory; Repeated game; Stochastic game; Subgame-perfect equilibrium; Competitive Markov decision process; Bounded rationality; Information; Prisoner's Dilemma; Cooperation

1. Introduction

It has long been known that repeated interaction can have major significance for the set of rational outcomes in a game theoretic scenario. For instance in the Prisoner's Dilemma, in which the Defect action is the only rational choice for both players in the one-shot game, the indefinitely repeated game has rational outcomes involving sustained cooperation (Luce & Raiffa, 1957). However, while determination of the set of Nash equilibria (NEs) of a one-shot matrix game is straightforward, much less is known about equilibria in repeated games. Here we focus primarily on one well-studied extension of NEs to repeated games, the subgame-perfect equilibrium (SPE; Selten, 1965). There are a number of strong results concerning the expected total rewards associated with the SPEs of a repeated game (Aumann & Shapley, 1994; Fudenberg & Maskin, 1986, 1991; Rubinstein, 1979, 1994), including a full characterization of the range of such payoffs in the iterated Prisoner's Dilemma (Stahl, 1991), but much less is known about the specific behaviors or policies involved. The difficulty is due to the fact that players in a repeated game can base actions on the full history of the game, so that the space of policies grows exponentially with the number of trials and reaches uncountable cardinality for a game of unbounded length, thus making identification of specific SPEs often quite complicated.

The multiplicity of SPEs in a repeated game also poses a difficulty for descriptive theory. The set of expected average payoffs achievable under SPEs includes all those from NEs of the one-shot game (because repetition of an NE constitutes an SPE), but the indeterminacy in the repeated game is far more severe than this. In fact, for every enforceable mixed outcome of the one-shot game (i.e., a reward vector in which

*Corresponding author. Department of Psychology, The University of Texas at Austin, 1 University Station A8000, Austin, TX 78712, USA. Fax: +1-512-471-5935.

E-mail address: mattj@psy.utexas.edu (M. Jones).

0022-2496/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2004.08.003

every player receives more than his or her minmax payoff), there is known to be an SPE of the repeated game with the same average payoffs (Aumann & Shapley, 1994; Fudenberg & Maskin, 1991). Thus the set of SPE payoffs forms a continuous region in joint reward space, spanning the full range of enforceable payoffs of the stage game. This and other so-called folk theorems would seem to render the SPE criterion essentially useless as a tool for predicting outcomes of real interactions.

From the cognitive perspective, there is the more general issue of how a player determines the best reply to arbitrary policies of his or her opponents. Given the uncountable class of policies available, it may appear that determination of the universally optimal solution is often extremely difficult. Thus one might expect that human players are forced to sacrifice optimality for efficiency, by using heuristics that base actions on incomplete or compressed information. However, it turns out that such a sacrifice is often unnecessary. Here we describe a broad class of situations in which universally optimal (or unboundedly rational) behavior can be achieved by an agent with incomplete information, bounded memory, and finite computational resources. Specifically, we consider situations in which all players choose actions based on some restricted subset of the information carried by the full history of the game, given by the value of a function I (which must satisfy two axioms, given below). For example, the information could consist of the outcomes of the previous n trials, or simply the total number of players who performed a particular action (such as Cooperate) on each trial without regard for who did what. A policy based on the information I is one that prescribes action probabilities based solely on the current value of I, regardless of other information about the game's history. Importantly, for the cases we consider, the value of I on each trial can be computed from the value on the previous trial along with knowledge of the outcome of that trial (see Definition 3.5). Therefore I represents the only information that must be retained from trial to trial.

It is proven here that whenever a player's opponents all use policies based on I, the player has a (universally) best reply that is also based solely on I. The information I can thus be thought of as a sufficient statistic, defined on the full history of the game, for determining optimal actions. Therefore any process by which the player can optimize his or her policy with respect to reliance on I (i.e., optimize within the class of I-based policies) will yield universally optimal (i.e., rational) behavior. Consequently, the restriction to finite memory and other types of "bounded rationality" are seen often to be unboundedly rational.

The main proof presented here relies on the concept of stochastic games, a generalization of repeated games in which the players' actions determine not only immediate payoffs but also the reward structure (i.e., the one-shot game to be played) on the next time step. We show that for any repeated or stochastic game, and any satisfactory choice of the function I, it is possible to construct an expanded stochastic game whose states correspond to the values of I. This construction also yields a payoff-preserving isomorphism between policies in the two games, under which I-based policies in the original game correspond to stationary policies (i.e., policies that depend only on the current state) in the expanded I-game. Furthermore, the preservation of payoffs implies that best replies and SPEs in the I-game correspond, respectively, to best replies and SPEs in the original game.

The next step in the proof relies on a result of Filar and Vrieze (1997, p. 173), stating that if all players in a stochastic game use stationary policies then no one has incentive to use a non-stationary policy. Applying this result to the expanded I-game yields a pair of "universality" results for the class of I-based policies in the original game. The first of these is the *best-reply universality theorem*, which states that if a player's opponents use I-based policies then the best reply within the class of I-based policies is a universally best reply. A corollary to this theorem is the *SPE universality theorem*, stating that restricted SPEs within the class of I-based policies are true SPEs within the full policy space. In other words, if all players are limited to I-based policies, and if every player is using a best reply to the others' policies subject to this constraint, then each player's policy is optimal in the full policy space and the SPE condition is satisfied. For example, if all players are restricted to n-back memory, that is they act based only on the outcomes of the previous n trials for some $n \geqslant 0$, then mutual optimization with respect to this constraint implies that each player's policy is universally optimal (i.e., with respect to policies of possibly unbounded memory). Thus from the perspective of each individual the restriction to n-back memory is no restriction at all, and determination of universally optimal behavior can be made with finite resources. Although the assumption that all players use exactly the same information, or have exactly the same memory span, may seem unrealistic, the critical point is that even if one player does incorporate additional information into his or her policy beyond that used by the opponents, that information cannot convey an advantage. Thus there is no incentive to expand one's memory beyond that of one's opponents.

Following derivation of the universality results, we present an analytical method for deriving all stationary SPEs of a stochastic game. The multiplicity of states in a stochastic game complicates the SPE criterion, even in the case of stationary policies, as the dependence of future states (and hence future rewards) on current actions can offset the equilibrium in immediate payoffs. However, these delayed effects can be accounted for by replacing the immediate payoffs in each state with the corresponding *outcome values* (denoted here by U), which include the expected value of all future rewards conditioned upon the policies to be used thereafter. Filar and Vrieze (1997, p. 220) show that a profile of stationary policies forms an SPE if and only if the strategies prescribed for each state form an NE of the one-shot U-game for that state. However, future rewards, and hence the values of U, depend on the policies being used, creating a complex circularity. Thus, although it is known that there exist SPEs consisting of stationary policies in any stochastic game (Fink, 1964), very little is known about the structure of these stationary SPEs (Filar & Vrieze, 1997, p. 230).

Here we outline a technique for handling the bi-directional dependence between policies and outcome values, and thus for deriving the stationary SPEs of the game, based on analysis of the qualitative form of the U-game associated with each state. By fixing certain inequalities among the components of U, the NEs of the U-games can be qualitatively determined. Each possible stationary SPE satisfying the assumptions placed on U then corresponds to a choice of one U-game NE for the strategy profile in every state. Once such a set of assignments has been made, the policies and U-values can be quantitatively determined. If the values of U thus obtained are consistent with the original qualitative assumptions (implying that the strategies in each state do in fact constitute U-game NEs), then the policy profile is an SPE. In summary, the set of stationary SPEs can be exhaustively determined by partitioning the possible values for U into a number of qualitative forms, and for each form testing all assignments of NEs to strategy profiles. This approach will likely be computationally unfeasible for more complex games, as the number of cases to be checked grows exponentially with the numbers of states, players, and actions, but as we show here in a case of four states and two players, each with two actions per state, it is reasonable for smaller games. When applied to the expanded I-game defined above, the method allows derivation of all SPEs in a repeated or stochastic game whose constituent policies base actions on some given compressed representation I of the game history. The class of SPEs potentially obtainable via this method is much broader than those based on stationary policies, and in particular includes all equilibria whose constituent policies can be implemented using finite-state automata.

The second half of this article applies the above results to the iterated Prisoner's Dilemma (IPD). One of the most well known policies in the IPD, which has been observed empirically both in humans (Rapoport & Chammah, 1965) and in animals (Wilkinson, 1984; Milinski, 1987) and has been successful in computer simulations (Axelrod, 1984) is Tit-For-Tat (TFT). In its idealized form, TFT bases actions only on the previous trial, by copying the opponent's last action. Because of the prevalence of TFT, as well as the general human bias towards recency effects in cognitive tasks (see Jones, 2003), we investigate here those policies in the IPD characterized by 1-back memory, meaning that players base their actions only on the outcome of the previous trial. We calculate two sets of SPEs within this restricted policy space: those that are deterministic (pure) and those that are symmetric (i.e., both players use the same policy relative to their roles in the game). The SPE universality theorem implies that all of these pairs constitute unrestricted SPEs, that is, individually fully rational outcomes. We find that TFT vs. TFT is an SPE but is unstable, in that it requires precise relationships among the game parameters. In addition there are five other pure SPEs (including all-defect). Two of these, including the Grim strategy (Friedman, 1971), are robust to variations in parameter values and will result in sustained mutual cooperation. In the case of mixed symmetric policy pairs there are ten SPEs (in addition to the symmetric pure ones). These equilibria all have the potential for manifesting varying degrees of cooperation.

2. Theoretical background

2.1. Matrix games

The basic element of our framework is the matrix game of von Neumann and Morgenstern (1944). In this game there are N players, with player k ($k \leqslant N$) having n_k actions. For ease of exposition we restrict to the situation $N = 2$; however, the central results (including all theorems presented in Section 3) carry over to the general case. The set of actions available to player k is denoted A_k. Probabilistic mixtures of these actions will be referred to as strategies, with the set of strategies for player k given by B_k. For any $b = \Sigma \alpha_i a_i \in B_k, b(a_i)$ denotes the probability α_i assigned to action a_i by strategy b. To each pair of actions $a_1 \in A_1, a_2 \in A_2$ is associated a reward for player k given by $r_k(a_1, a_2)$. For ease of notation, the expected reward $E_{b_1,b_2}[r_k(a_1,a_2)] = \Sigma_{a_1 \in A_1, a_2 \in A_2} b_1(a_1) b_2(a_2) r_k(a_1, a_2)$ is denoted $r_k(b_1, b_2)$. All rewards are assumed to be in units of players' utilities, and completely capture the players' preferences among outcomes (e.g., the opponent's reward is irrelevant).

For example, the Prisoner's Dilemma can be characterized by the following payoffs:

$$[(r_1(a_1,a_2), r_2(a_1,a_2))]_{a_1 \in A_1, a_2 \in A_2} = \begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{bmatrix} (1,1) & (x,y) \\ (y,x) & (0,0) \end{bmatrix} \end{array}. \quad (1)$$

Here C and D represent the Cooperate and Defect actions. Thus for example when player 1 (row player) cooperates and player 2 defects, their respective payoffs are x and y (upper right entry on RHS of Eq. (1)). The parameter y is the "temptation payoff" and is taken to be greater than 1; the "sucker's payoff" x is always negative. Because utilities are only defined in terms of their implications for preferences among lotteries (von Neumann & Morgenstern, 1944), they are measured on an interval scale, and thus the payoffs can be normalized so that both players receive rewards of 1 when both cooperate and 0 when both defect.

A key solution concept in matrix games is the *Nash equilibrium* (NE; Nash, 1950). An NE consists of a pair of strategies b_1^* and b_2^* that are best replies to each other, in the sense that neither player can improve his or her expected reward by a unilateral change in strategy:

$$\begin{aligned}&\forall b_1 \in B_1 : r_1(b_1, b_2^*) \leqslant r_1(b_1^*, b_2^*),\\ &\forall b_2 \in B_2 : r_2(b_1^*, b_2) \leqslant r_2(b_1^*, b_2^*).\end{aligned} \tag{2}$$

Because of this definition, an NE is often considered an individually rational outcome for both players. As can be readily seen from the payoff matrix of the Prisoner's Dilemma (Eq. (1)), the only NE of that game is (D, D). Indeed, D is the dominant action for both players, in the sense that each player will fare better by defecting than cooperating regardless of the opponent's action.

2.2. Repeated games

A repeated game models the situation in which a (one-shot) matrix game is played multiple times between the same players. In such a game the players are able to choose actions on each trial contingent on the outcomes of earlier trials. Formally, a *policy* in a repeated game is a function yielding a strategy (for the one-shot game) at every stage as a function of the prior history. Thus if we define the set of histories as

$$\begin{aligned}H = \{&(a_1^0, a_2^0, a_1^1, a_2^1, \ldots, a_1^{t-1}, a_2^{t-1}) | t \geqslant 0,\\ &\forall k \forall \tau < t [a_k^\tau \in A_k]\}\end{aligned} \tag{3}$$

(where $t = 0$ corresponds to the initial history $h = \emptyset$) then a policy for player k can be written as

$$f : H \to B_k. \tag{4}$$

In what follows, the history h is written in square brackets, with $f[h](a)$ giving the probability assigned by f to action a following history h.

With policies thus defined, the concept of NE can be carried directly over to repeated games, as a pair of policies that are optimal with respect to unilateral deviation. However, this definition is deficient as a criterion for rational outcomes because it places no constraints on the strategies players would use following histories that cannot occur under the given policies. These strategies do affect the NE criterion, as they enter into players' evaluations of alternative policies (see, e.g., Osborne & Rubinstein, 1994, pp. 95–96). In order to account for this concern, the NE is generalized to the *subgame-perfect equilibrium* (SPE; Selten, 1965). Two policies form an SPE if and only if they give rise to an NE in the sub-game obtained by starting play at any arbitrary history. Formally, for any policy f and any history h, let f^h denote the policy derived from f by appending h to the beginning of any history:

$$f^h[h'] = f[(h, h')], \tag{5}$$

where (h, h') is the concatenation of h and h'. Then f_1 and f_2 form an SPE iff (f_1^h, f_2^h) is an NE for every h.

In a fixed length repeated game, the constituent one-shot game is played for T trials, with T a fixed value known to both players in advance. In such a situation the set of SPEs can be calculated straightforwardly via the well-known backward induction technique, in which optimal actions are determined first for the final trial (as NEs of the one-shot game) and then for earlier trials by adding to the immediate payoffs the expected future payoffs determined in the previous steps of the induction (see, e.g., Osborne & Rubinstein, 1994, pp. 99–100). In the case of the IPD, the backward induction argument implies that the only SPE is the one in which both players defect on every trial. This is because the only NE for the final trial is (D, D), and on each previous trial the effective payoffs are the same as in the one-shot game (original payoffs plus 0), under the inductive assumption that each player will defect on all following trials (Luce & Raiffa, 1957).

When the game length is infinite (or finite but stochastic) backward induction breaks down, as there is no final trial on which to anchor the argument. In order to analyze this situation, however, a criterion is needed for evaluating a reward sequence of unbounded length. The criterion considered here is the *total discounted reward*:[1]

$$V((r^t)_{t \geqslant 0}) = \sum_{t \geqslant 0} r^t \gamma^t. \tag{6}$$

Here γ is the *discount factor*, with $0 \leqslant \gamma < 1$. This sum is guaranteed to exist, assuming a bounded range of payoffs for the constituent game. Total discounted reward is also a useful criterion because it corresponds to the expected total (non-discounted) reward under a finite stochastic-length game with constant continuation probability γ, that is a game of length T with $P[T \geqslant t] = \gamma^{t-1}$. This correspondence allows the more realistic stochastic-length game to be modeled using the math-

[1] Another commonly used criterion is the *limiting average*: $V_{\lim}((r^t)_{t \geqslant 0}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} r^\tau$. The universality results presented in the following section (Theorems 3.3 and 3.4, as well as Lemma 2.2 and Corollary 3.1) hold for this criterion as well.

ematically more convenient infinite-length discounted game.

Given choices of policies f_1 and f_2 for the players, we denote the expectation of the total discounted reward for player k as

$$V_k(f_1,f_2) = E\left[\sum_{t\geqslant 0} r_k^t \gamma^t | f_1, f_2\right]. \tag{7}$$

Similarly, the expected total discounted reward starting in some history h, with length τ, is given by

$$\begin{aligned} V_k(f_1,f_2,h) &= E\left[\sum_{t\geqslant\tau} r_k^t \gamma^{t-\tau} | h, f_1, f_2\right] \\ &= E\left[\sum_{t\geqslant 0} r_k^t \gamma^t | f_1^h, f_2^h\right] = V_k(f_1^h, f_2^h), \end{aligned} \tag{8}$$

where the first expectation is conditioned upon the game following h on trials 0 through $\tau - 1$ and evolving according to f_1 and f_2 thereafter.

In keeping with the definition of SPE, we use the term *best reply*, in the context of policies, to mean a policy that is optimal, given that of the opponent(s), following any history:

Definition 2.1. A policy f_1^* is a best reply to f_2 iff

$$\forall h \forall f_1: \quad V_1(f_1,f_2,h) \leqslant V_1(f_1^*,f_2,h). \tag{9}$$

2.3. Stochastic games

An extension to the repeated game is the *stochastic game*, in which the one-shot game differs from trial to trial (see Filar & Vrieze, 1997, for a thorough introduction). Each one-shot game corresponds to a *state* of the stochastic game, with transition probabilities among the states dependent on the players' actions. Thus a stochastic game can be characterized by the set of states S (assumed here to be countable), an initial state s_0, the set of actions $A_k(s)$ available to player k in state s, immediate rewards $r_k(s, a_1, a_2)$ for $a_j \in A_j(s)$ for any s, and transition probabilities $p(s'|s, a_1, a_2)$ giving the probability that state s' will follow s given actions a_1 and a_2. As before, $B_k(s)$ will denote the space of formal convex mixtures of elements of $A_k(s)$, that is, the set of mixed strategies available to player k in state s. When S consists of a single state, the stochastic game reduces to a repeated game.

In a stochastic game policies can take as input past states as well as actions, and thus the history must be expanded to include the sequence of past states, as well as the present state. For technical reasons, the set H of histories is restricted to those that are realizable given the transition probabilities of the game, that is sequences $(s^0, a_1^0, a_2^0, \ldots, s^{t-1}, a_1^{t-1}, a_2^{t-1}, s^t)$ satisfying $p(s^{\tau+1}|s^\tau, a_1^\tau, a_2^\tau) > 0$ for all $\tau < t$.[2]

An important class of policies in a stochastic game is the class Ω of stationary policies. A stationary policy is one in which strategies depend only on the present state. Equivalently, a stationary policy f is one that can be written as $\tilde{f} \circ c$, where $\tilde{f}$ is a map from states to action probabilities: $s \mapsto b \in B_k(s)$, and the function c: $H \to S$ returns the terminal (i.e., current) state of any history. Often a stationary policy f is identified with $\tilde{f}$, in that f $[s](a)$ represents the probability of choosing action a in state s (i.e., following any history ending in s).

When both players use stationary policies, the expected future rewards for both players depend only on the present state (rather than on the full history), and we can define the *state value*:

$$\begin{aligned} V_k(f_1,f_2,s) &= E\left[\sum_{t\geqslant\tau} r_k^t \gamma^{t-\tau} | s^\tau = s, f_1, f_2\right] \\ &= V_k(f_1,f_2,h) \forall h \ni c(h) = s. \end{aligned} \tag{10}$$

Further, in order to account for the delayed rewards associated with actions due to their effects on subsequent states and strategies, define the *outcome value*:

$$\begin{aligned} U_k(f_1,f_2,s,a_1,a_2) &= E\left[\sum_{t\geqslant\tau} r_k^t \gamma^{t-\tau} | s^\tau = s, a_1^\tau = a_1, a_2^\tau = a_2, f_1, f_2\right] \\ &= r_k(s,a_1,a_2) + \gamma \sum_{s'} \left[V_k(f_1,f_2,s') p(s'|s,a_1,a_2)\right]. \end{aligned} \tag{11}$$

The third expression here gives U explicitly as a sum of immediate and delayed rewards, and shows its dependence on V.[3] Furthermore, the value V associated with any state can readily be seen to equal the expected outcome value, conditioned upon the strategies used in that state:

$$V_k(f_1,f_2,s) = \sum_{a_1,a_2} \left[U_k(f_1,f_2,s,a_1,a_2) f_1[s](a_1) f_2[s](a_2)\right]. \tag{12}$$

Often the policies are suppressed and we write $V_k(s)$ and $U_k(s, a_1, a_2)$.

One interpretation of U is as the value of a (joint) 1-step deviation, that is, the expected total discounted reward when the players choose arbitrary actions a_1 and a_2 on the first (or current) trial and then follow their policies thereafter. The concept of 1-step deviation is especially useful as it provides a necessary and sufficient criterion for SPE: A pair of policies forms an SPE if and only if neither player can gain an advantage via

[2] Later, when we introduce games with constraints on the starting state s^0, H will be accordingly further restricted.

[3] The outcome values are nearly the same as the action values used in Q-learning (Watkins, 1989), except that they depend on the joint actions of both players (see, e.g., Hu & Wellman, 1998).

unilateral 1-step deviation following any history (e.g., Osborne & Rubinstein, 1994, p. 153). Further, when both players' policies are stationary, it is sufficient to check this condition in each state rather than following every possible history. Therefore, in the case of stationary policies, a necessary and sufficient condition for SPE is that in every state each player uses a strategy that maximizes the expected value of U, as conditioned upon the opponent's strategy in that state (Filar & Vrieze, 1997, p. 220). This fact is formalized in the following lemma. Later we demonstrate how this result allows explicit calculation of stationary SPEs in a stochastic game.

Lemma 2.1. (Filar & Vrieze, 1997, p. 220). *A pair of stationary policies f_1 and f_2 forms an SPE if and only if for every state s the strategies $f_1[s]$ and $f_2[s]$ constitute an NE for the one-shot "U-game" with payoffs given by*

$$\left[\left(U_1(f_1, f_2, s, a_1, a_2), U_2(f_1, f_2, s, a_1, a_2)\right)\right]_{a_1 \in A_1(s),\ a_2 \in A_2(s)}. \tag{13}$$

In the case of a repeated game, the stationary policies are just those that give the same strategy on every trial. Because these policies give no dependence of future events (states or strategies) upon present actions, the components of U only differ from the immediate payoff matrix by a constant, and thus the U-game and the original one-shot game have the same NEs (see Eq. (11)). Thus in this case Lemma 2.1 reduces to the statement made earlier that repetition of a single-trial NE constitutes an SPE. One contribution of the results presented here (see Theorem 3.5) is to extend the applicability of Lemma 2.1 beyond stationary policies to policies that depend on past states and actions. This is a particular improvement in the case of repeated games, as it allows players' behavior to depend on each other's past actions, which in turn forces players to consider the effects of their actions on their opponents' future strategies. As is shown here in the case of the Prisoner's Dilemma, this dependence has significant consequences for the types of rational behaviors that can arise.

2.4. Competitive Markov decision processes

A stochastic game can also be cast as a Markov decision process (MDP) with multiple agents (Filar & Vrieze, 1997). In a standard MDP, there is a single agent interacting in an environment with a discrete set of states. Associated with each state is a set of possible actions, which determine the player's immediate payoff as well as the probability distribution over the state that will obtain on the following time step. A stochastic game is therefore a competitive MDP, in which rewards and transition probabilities are jointly determined by the simultaneous actions of multiple agents. Furthermore, a stationary policy for all but one of the players induces a (standard) MDP for the remaining player, with transition probabilities and expected rewards at each state obtained by conditioning upon the action probabilities dictated by the stationary policies of the opponent(s) (Filar & Vrieze, 1997, p. 172). A best reply for the player is therefore given by any optimal policy in the MDP.[4] Now it is well known that the set of optimal policies in an MDP includes a pure stationary policy (assuming a finite and bounded number of actions per state; Blackwell, 1965). The conclusion is summarized in the following lemma.

Lemma 2.2. (Filar & Vrieze, 1997, p. 173). *If all of a player's opponents in a stochastic game use stationary policies, then the set of best replies for the player includes a pure stationary policy.*

3. Universality of bounded policy classes

Lemma 2.2 implies that if a player's opponents use (possibly pure) stationary policies then the player has no incentive but to do the same. Using a more complex history-dependent policy cannot increase the player's expected payoff. This is summarized in the following *universality* property as applied to the classes of stationary and pure stationary policies.

Definition 3.1. A policy class R is *best-reply universal* if, whenever all of a player's opponents follow policies from R, the player's set of best replies includes a member of R.

A further consequence of Lemma 2.2 is the following:

Corollary 3.1. *If all players in a stochastic game use (pure) stationary policies and each player's policy is optimal (in the sense of* Definition 2.1*) with respect to the others' policies subject to the (pure) stationarity constraint, then the policy profile forms an SPE.*

Proof. Lemma 2.2 implies that each player has a universal best reply that is pure and stationary. Therefore optimizing subject to the (pure) stationarity constraint yields a universal best reply. Since every player is using a best reply to his or her opponents, the SPE condition is satisfied. □

Corollary 3.1 states that any policy profile that satisfies the SPE condition under the restriction to (pure) stationary policies is a true SPE. This notion is formalized with the following two definitions.

[4] The set of policies in the stochastic game is strictly larger than that for the MDP, because the former can take as input the opponents' actions in addition to the states and rewards they resulted in. However, when the opponents' policies are stationary, this additional information cannot be used to improve expected rewards (Filar & Vrieze, 1997, p. 168).

Definition 3.2. A *restricted SPE* within a policy class R is a profile of policies in R that are best replies to each other from among the policies of R. In the case of two players, $f_1^*, f_2^* \in R$ form a restricted SPE within R iff the following conditions hold:

$$\begin{aligned}&\forall h, \forall f_1 \in R: V_1(f_1, f_2^*, h) \leqslant V_1(f_1^*, f_2^*, h),\\ &\forall h, \forall f_2 \in R: V_2(f_1^*, f_2, h) \leqslant V_2(f_1^*, f_2^*, h).\end{aligned} \tag{14}$$

Definition 3.3. A policy class R is *SPE universal* if every restricted SPE within R is an unrestricted SPE, that is an SPE within the space of all (mixed, history-dependent) policies.

Corollary 3.1 implies that the classes Ω of stationary policies and $\bar{\Omega}$ of pure stationary policies are both SPE universal. As can be seen from the proof, any class that is best-reply universal is also SPE universal. SPE universality of a policy class R implies that boundedly rational outcomes subject to the restriction to R are in fact fully rational, as no player could benefit by switching to a policy outside of R.

We now show that these universality properties apply in a wide range of cases beyond the classes Ω and $\bar{\Omega}$. Specifically, we address situations where all players choose actions based on the same restricted subset of the information carried by the full history of the game. This information will be denoted I. More precisely, let I be any function on histories, that is a mapping $H \rightarrow \Xi$, where Ξ is an arbitrary set acting as the image space of I. (It will be seen shortly that Ξ also represents the set of states in an expanded stochastic game determined by I.) Now let F_I be the class of policies that determine action probabilities based solely on the value of I, and let $\bar{F}_I$ be the subclass of pure policies within F_I. Analogous to the definition of stationary policies, f is a policy for player k in F_I if and only if there exists a map $\hat{f}: \Xi \rightarrow \cup_s B_k(s)$ such that $f = \hat{f} \circ I$. Often f is identified with $\hat{f}$, with $f[\xi]$ representing $f[h]$ for any h satisfying $I(h) = \xi$. This formalizes the notion of acting based solely on the information carried by I. Note that when $I \equiv h$ (the identity function) we have $F_I = F$, the full space of policies; likewise $F_c = \Omega$ and $\bar{F}_c = \bar{\Omega}$ (recall that c returns the current state of any history).

We show that F_I and $\bar{F}_I$ are best-reply and SPE universal provided I satisfies the following two axioms:

Definition 3.4. The function I is *sufficient* if it gives unambiguous knowledge of the present state, that is I determines a well-defined mapping Φ: $\Xi \rightarrow S$ satisfying $\Phi \circ I = c$.

Definition 3.5. The function I is *deterministic* if the present value of I, the actions of the two players, and the ensuing state are sufficient to determine the new value of I. More precisely, there must exist a map

$$\Psi : \Xi \times \cup_s A_1(s) \times \cup_s A_2(s) \times S \rightarrow \Xi$$

satisfying

$$\Psi(I(h), a_1, a_2, s) = I([h, a_1, a_2, s]). \tag{15}$$

for all h, s, $a_1 \in A_1(c(h))$, and $a_2 \in A_2(c(h))$ satisfying $p(s|c(h), a_1, a_2) > 0$. Here $[h, a_1, a_2, s]$ represents the new history obtained from h by appending actions a_1 and a_2 and subsequent state s. (Note that the present definition coincides with the definition of determinism in automata theory, under the interpretation of I as an automaton with states indexed by Ξ and input given by the triple $[a_1^t, a_2^t, s^{t+1}]$.)

The identity function h and the current-state function c both satisfy these two axioms. Other examples include n-back memory, where I encodes the states and actions of the previous n trials along with the current state, and n-back state memory, which encodes only the states of the last n trials (including the current one). Also, for many-player games in which all players always have the same set of actions, I can encode simply the number of players who performed each action (e.g., Cooperate) on each of the last n trials.

The significance of the sufficiency and determinism axioms is that they allow the definition of an expanded game in which each state of the original game is split into substates corresponding to the possible values of I compatible with that state.

Lemma 3.2. *Let Γ be a stochastic game and let I: $H \rightarrow \Xi$ be any sufficient and deterministic function on histories. Define the stochastic process $\xi^t = I(s^t)$, where $(s^t)_{t \geqslant 0}$ is the sequence of states arising in Γ (conditioned on the policies f_1 and f_2). Then there exists a new stochastic game Γ_I with state-space Ξ whose policies correspond one-to-one with those of Γ in a manner that preserves expected payoffs to both players, with the sequence of resulting states and rewards given by the process $(\xi^t, r_1^t, r_2^t)_{t \geqslant 0}$.*

Proof. The proof proceeds constructively, by explicitly defining the actions, rewards, and transition probabilities for Γ_I and then verifying that the resulting game has the stated correspondences with Γ. Because I is assumed to be sufficient, we can define Γ_I by splitting each state s of Γ into states $\xi \in \Phi^{-1}(s)$. Actions and rewards can be carried directly over via $A_k(\xi) = A_k(\Phi(\xi))$ and $r_k(\xi, a_1, a_2) = r_k(\Phi(\xi), a_1, a_2)$. Using determinism of I we can define transition probabilities by

$$p(\xi'|\xi, a_1, a_2) = \begin{cases} p(\Phi(\xi')|\Phi(\xi), a_1, a_2), & \Psi(\xi, a_1, a_2, \Phi(\xi')) = \xi', \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

Thus transition probabilities in Γ_I are nearly the same as in Γ, with the constraint that the succeeding Ξ-state ξ'

must correspond to the value of I dictated by the prior Ξ-state, the actions, and the succeeding S-state $\Phi(\xi')$. In addition, the set of valid starting states must be restricted to the possible values of I evaluated on length-0 histories (i.e., histories consisting of just a starting state in Γ), or equivalently those ξ satisfying $I((\Phi(\xi))) = \xi$.

Observe now that there exists a bijection between realizable histories in the two games, given by

$$\begin{aligned}\Theta : &(s^0, a_1^0, a_2^0, \ldots, s^{t-1}, a_1^{t-1}, a_2^{t-1}, s^t) \\ &\mapsto (\xi^0, a_1^0, a_2^0, \ldots, \xi^{t-1}, a_1^{t-1}, a_2^{t-1}, \xi^t)\end{aligned} \tag{17}$$

with $\xi^\tau = I(s^0, a_1^0, a_2^0, \ldots, s^{\tau-1}, a_1^{\tau-1}, a^{\tau-1}, s^\tau)$. Sufficiency of I guarantees that Θ is injective, as Φ allows us to define a left inverse to Θ via

$$\begin{aligned}&(\xi^0, a_1^0, a_2^0, \ldots, \xi^{t-1}, a_1^{t-1}, a_2^{t-1}, \xi^t) \\ &\mapsto (\Phi(\xi^0), a_1^0, a_2^0, \ldots, \Phi(\xi^{t-1}), a_1^{t-1}, a_2^{t-1}, \Phi(\xi^t)).\end{aligned} \tag{18}$$

Surjectivity of Θ is a result of determinism of I, and the fact that the transition probabilities defined for Γ_I make the set H_{Γ_I} of realizable histories in the expanded game correspond precisely to the image of Θ. (Recall that H_{Γ_I} is defined as the set of histories that are realizable given the game's transition probabilities and eligible starting states.) Indeed, the conditions

$$I((\Phi(\xi^0))) = \xi \text{ and } \forall \tau < t : \Psi(\xi^\tau, a_1^\tau, a_2^\tau, \Phi(\xi^{\tau+1})) = \xi^{\tau+1} \tag{19}$$

are necessary and jointly sufficient both for the history $(\xi^0, a_1^0, a_2^0, \ldots, \xi^{t-1}, a_1^{t-1}, a_2^{t-1}, \xi^t)$ to be realizable and for it to be in the image of Θ.

Finally, note that Θ also implies a bijection between policies of the two games, via $f \mapsto f \circ \Theta^{-1}$ (where f is any policy in Γ). It is now straightforward to verify (by induction on t) that the probability distributions on the sequences of I-values $(I(s^t))_{t \geqslant 0}$ and rewards $(r_k^t)_{t \geqslant 0}$ implied by arbitrary policies f_1 and f_2 in Γ are identical to those on Ξ-states $(\xi^t)_{t \geqslant 0}$ and rewards $(r_k^t)_{t \geqslant 0}$ implied by the corresponding policies $f_1 \circ \Theta^{-1}$ and $f_2 \circ \Theta^{-1}$ in Γ_I. □

The expanded game Γ_I allows the results of Lemmas 2.1 and 2.2 and Corollary 3.1 to be extended from stationary policies to those based on the information I. First, note that because the isomorphism between policies in the two games preserves payoffs, it also preserves both best replies and SPEs. Therefore application of the earlier results to Γ_I leads to novel conclusions regarding best replies and SPEs in Γ.

Theorem 3.3. (Best-Reply Universality). *If I is sufficient and deterministic then F_I and $\bar{F}_I$ are best-reply universal.*

Proof. Observe that F_I and $\bar{F}_I$ correspond precisely to the classes of stationary and pure stationary policies in the expanded game Γ_I. Indeed, a policy f for player k in Γ is an element of F_I if and only if there exists a map $\hat{f} : \Xi \to \cup_s B_k(s)$ with $f = \hat{f} \circ I$, whereas the corresponding policy $f \circ \Theta^{-1}$ for Γ_I is stationary if and only if there exists a map $\tilde{f} : \Xi \to \cup_\xi B_k(\xi)$ with $f \circ \Theta^{-1} = \tilde{f} \circ c$, or equivalently, $f = \tilde{f} \circ c \circ \Theta$. Using the equivalence $\cup_s B_k(s) = \cup_\xi B_k(\xi)$ and the functional relation $c \circ \Theta = I$, the maps $\tilde{f}$ and $\hat{f}$ are seen to be identical, and thus the existence of one implies that of the other. The argument for pure policies is similar, with B_k replaced by A_k.

Now let f be any member of F_I (respectively $\bar{F}_I$). Because $f \circ \Theta^{-1}$ is (pure) stationary in Γ_I, there exists a (pure) stationary best reply for the opponent by Lemma 2.2. The policy in Γ that corresponds to this best reply lies in F_I ($\bar{F}_I$) and is a best reply to f. Therefore F_I ($\bar{F}_I$) is best-reply universal. □

Theorem 3.4. (SPE Universality). *If I is sufficient and deterministic then F_I and $\bar{F}_I$ are SPE universal.*

Proof. The correspondence between F_I (respectively $\bar{F}_I$) and Ω_{Γ_I} ($\bar{\Omega}_{\Gamma_I}$) implies that for any strategy pair (f_1, f_2) that forms a restricted SPE within F_I ($\bar{F}_I$), the corresponding pair $(f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1})$ is an SPE within Ω_{Γ_I} ($\bar{\Omega}_{\Gamma_I}$). Now, $\Omega_{\Gamma_I}(\bar{\Omega}_{\Gamma_I})$ is SPE universal in Γ_I by Lemma 3.1, and therefore $(f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1})$ is an SPE within the full policy space F_{Γ_I}. Using the bijection Θ once more we see that (f_1, f_2) is an unrestricted SPE for Γ. Therefore F_I and $\bar{F}_I$ are SPE universal. □

Finally, the outcome-value criterion for stationary SPEs given in Lemma 2.1 can be extended to SPEs in I-based policies, using the outcome values U^I of Γ_I.

Theorem 3.5. *Assume I is sufficient and deterministic. A pair of policies f_1, $f_2 \in F_I$ forms an SPE if and only if for every value ξ of I the strategies $f_1[\xi]$ and $f_2[\xi]$ constitute an NE for the one-shot U^I-game where payoffs are given by*

$$\begin{aligned}&\big[\big(U_1^I(f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1}, \xi, a_1, a_2), \\ &U_2^I(f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1}, \xi, a_1, a_2)\big)\big]_{a_1 \in A_1(\Phi(\xi)), a_2 \in A_2(\Phi(\xi))}.\end{aligned} \tag{20}$$

Proof. Apply Lemma 2.1 to the policies $f_1 \circ \Theta^{-1}$ and $f_2 \circ \Theta^{-1}$ in Γ_I. Note that $f_1[\xi] = f_1 \circ \Theta^{-1}[\xi]$ and $f_2[\xi] = f_2 \circ \Theta^{-1}[\xi]$. □

4. Bounded memory in the iterated Prisoner's Dilemma

The Prisoner's Dilemma has been a focal point of game theory research because it embodies a conflict between individual and collective rationality—the outcome that results from each player choosing what is best for him or her leaves both players worse off. In other words, the game's unique NE is Pareto inefficient. It has long been known that a rational basis for cooperation requires that the same players engage in the game repeatedly with no predetermined final trial (Luce & Raiffa, 1957). In this situation the full set of rational outcomes, in terms of expected payoffs associated with SPEs, is known, as a function of the discount factor (Stahl, 1991). However, little is known about the SPE policies themselves.

Here we apply the results of the previous section to the indefinitely repeated Prisoner's Dilemma for the case of 1-back policies, that is policies that determine strategies based only on the outcome of the previous trial. Thus we consider the function I that can take on four values, denoted CC, CD, DC, and DD (with player 1's action given first), according to the most recent action pair in any history.[5] The information I is clearly deterministic, as it is fully determined by the most recent actions, and is trivially sufficient since the IPD is a repeated game (i.e., there is only one state). We can therefore define the expanded game Γ_I, which has four states ξ corresponding to the values of I, with transition probabilities dependent only on the actions: $\forall\xi\ p(CC|\xi, C, C) = 1$, et cetera. Stationary policies in this game are functions that specify a probability of cooperating in each state (i.e., following each possible outcome), and will be denoted $f = \begin{bmatrix} f[CC] & f[CD] \\ f[DC] & f[DD] \end{bmatrix}$, with $f[\xi]$ indicating the probability of cooperating in state ξ.

The SPE universality theorem implies that any pair of policies that base replies only on the outcome of the previous trial (corresponding to stationary policies in Γ_I), and that are best replies to each other with respect to this constraint, form an SPE for the IPD. Here we explicitly calculate all SPEs of this type from among two classes: those involving pure policies and those that are symmetric (in that both players use the same policies, relative to their roles in the game). In the first case we determine the set of best replies to all 16 pure stationary policies, from among that same set, and check for pairs that are mutual best replies. In the case of mixed policies, the analysis relies on the result of Theorem 3.5 that any pair of I-based policies forming an SPE must jointly prescribe a U^I-game NE in every state. The nature of the state transitions in Γ_I implies that the values of U^I depend only on present actions and not on the state, and thus there is only one U-game to consider. (Henceforth the superscript I is omitted in writing U; U will be understood to represent the outcome values in Γ_I.) Knowing the qualitative form of this U-game, and thus the qualitative nature of its NEs, strongly constrains the set of possible SPEs. Thus the approach taken is to partition the (4-dimensional) space of possible U values into a number of regions each associated with a fixed "U-shape," and to systematically determine the set of symmetric SPEs for each region. (Focusing on symmetric SPEs allows us to assume the U-game is symmetric, thus simplifying the search process. However, the general approach applies to asymmetric SPEs as well.)

4.1. Pure policies

The approach to the case of pure policies is to determine the best pure reply to each policy and to look for matches (i.e., mutual best replies). The SPE universality theorem as applied to pure policies ensures that any such match will be an unrestricted SPE. These deterministic SPEs form boundary points of the SPE manifold in 8-dimensional joint 1-back policy space.

Because of the subgame-perfect criterion, a best reply for player 1 to a policy of player 2 is one that simultaneously maximizes $V_1(\xi)$ for all four values of ξ (see Definition 2.1). An example calculation, for the case $f_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, is given in Appendix A. The results of all 16 calculations are displayed in Table 1, which shows the set of best pure stationary replies to each of the 16 pure stationary policies in Γ_I. Best-reply universality implies that each of these is a universally best reply, without restriction to 1-back memory. From Table 1 it can be seen that, for sufficiently small γ, the best reply to any policy is the all-defect policy $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. This is because when future rewards become sufficiently insignificant, the immediate incentive to defect, due to dominance of that action in the single-trial game, overpowers any long-term considerations. Furthermore, for the majority of policies, the best reply is all-defect independent of the value of γ. However, there are many policies for which the best reply can involve cooperation, implying that if the opponent were using such a policy then it would be rationally justified to cooperate in certain states.

The tabulation of best replies now allows determination of the set of pure stationary SPEs for Γ_I (equivalently the 1-back pure SPEs for the IPD). These SPEs are pairs of policies (f, g) such that f is a best reply to g and g^{T} is a best reply to f^{T} (where T denotes the transpose operator, which corresponds to switching between players under the matrix representation used

[5] This definition does not determine the value of $I(\varnothing)$, which can be taken to be any of the four values. Because the SPE requirement applies following all histories (or in every state, for stationary policies), this choice is irrelevant. An alternative is to have I take on a fifth value on the initial trial. Because this state only occurs prior to the other four, its addition would not affect the analyses presented here. Furthermore, the U-game for the fifth state would match that of the other four states, and thus every SPE for the 5-state model is given by an SPE for the 4-state model together with a choice of a U-game NE for play on the initial trial.

Table 1
Best replies to all pure 1-back policies in the IPD

Policy (player 2)	Best replies (player 1)	Discounting range
$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$	$-\frac{x}{y} \leqslant \gamma < 1$
	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$	$\gamma = -\frac{x}{y}$
	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$0 \leqslant \gamma \leqslant -\frac{x}{y}$
$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$-\frac{x}{y-x} \leqslant \gamma < 1$
	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\gamma = -\frac{x}{y-x}$
	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$0 \leqslant \gamma \leqslant -\frac{x}{y-x}$
$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$1 - \frac{1}{y} \leqslant \gamma < 1$
	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$0 \leqslant \gamma \leqslant 1 - \frac{1}{y}$
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$y - 1 \leqslant \gamma < 1$
	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\gamma = y - 1$
	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$0 \leqslant \gamma \leqslant y - 1$
$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	All γ
$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\max\{\frac{y-1}{1-x}, -\frac{x}{1-x}\} \leqslant \gamma < 1$
	$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$	$\gamma = -\frac{x}{1-x}$
	$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$	$\gamma = \frac{y-1}{1-x}$
	$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$1 - \frac{1}{y} \leqslant \gamma \leqslant -\frac{x}{1-x}$
	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$	$\gamma = -\frac{x}{y}$, with $x + y = 1$
	$\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$	$-\frac{x}{y} \leqslant \gamma \leqslant \frac{y-1}{1-x}$

Table 1 (*continued*)

Policy (player 2)	Best replies (player 1)	Discounting range
	$\begin{bmatrix}0&0\\0&1\end{bmatrix}\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\gamma = -\frac{x}{y}$
	$\begin{bmatrix}0&1\\0&0\end{bmatrix}\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\gamma = 1-\frac{1}{y}$
	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$0\leqslant\gamma\leqslant \min\left\{1-\frac{1}{y}, -\frac{x}{y}\right\}$
$\begin{bmatrix}1&1\\0&1\end{bmatrix}$	$\begin{bmatrix}1&1\\0&1\end{bmatrix}$	$y-1\leqslant\gamma<1$
	$\begin{bmatrix}1&0\\0&0\end{bmatrix}\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}1&1\\0&0\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}\begin{bmatrix}0&1\\0&0\end{bmatrix}\begin{bmatrix}0&1\\0&1\end{bmatrix}$	$\gamma = y-1$
	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$0\leqslant\gamma\leqslant y-1$
$\begin{bmatrix}1&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&1\end{bmatrix}$	$-\frac{x}{y-x}\leqslant\gamma<1$
	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$0\leqslant\gamma\leqslant -\frac{x}{y-x}$
$\begin{bmatrix}1&1\\1&1\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	All γ

Notes: Discounting range gives constraints on γ for the given reply(s) to be optimal. The ranges for γ depend on the payoff parameters x and y; not all cases will occur for all values of these parameters.

here). In addition to the all-defect–all-defect pair, there are five other such pairs, all of which include the potential for cooperation (see Table 2). The first of these is the Grim policy (Friedman, 1971), which can cooperate indefinitely but responds to any deviation with unending defection. Equilibrium #2 involves a similar policy, with the crucial difference that both players will reconcile following mutual defection, thus ensuring that the pattern of behavior will always settle quickly (i.e., within 2 trials) into sustained mutual cooperation. Because the guarantee of reconciliation reduces the punishment for defection, the discount factor must be somewhat greater to support this equilibrium than is required for Grim. Equilibrium #3 leads either to permanent mutual defection or to an alternating sequence of one player cooperating and the other defecting. (Note that the requirement $\gamma<1$ implies this equilibrium can only occur if the joint reward associated with the *CD* outcome is greater than that for *DD*, i.e. $x+y>0$.) This is a very tenuous equilibrium, in that it only occurs for a precise value of γ; we will see, however, that there is a family of SPEs in mixed policy space that jointly exist for a full range of γ-values, of which the present SPE is a boundary point. SPE #4 is a minor variation of #3, and is the only asymmetric equilibrium of the group (only one version is displayed in Table 2). In the final cooperative SPE, under the given precise values of the payoffs and discount factor, the TFT policy makes the other player's entire policy irrelevant to his or her aggregate reward (a generalization of this situation that is robust to parameter values will be seen in the next section). This may be seen as an analogue of the maxmin NE that occurs in certain zero-sum one-shot games.

Table 2
Pure 1-back SPEs for the IPD

Equilibrium policies		
Player 1	Player 2	Discounting range
$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	$1-\frac{1}{y}\leqslant\gamma<1$
$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$y-1\leqslant\gamma<1$
$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}0&1\\0&0\end{bmatrix}$	$\gamma=-\frac{x}{y}$
$\begin{bmatrix}0&0\\1&0\end{bmatrix}$	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\gamma=-\frac{x}{y}$
$\begin{bmatrix}1&0\\1&0\end{bmatrix}$	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\gamma=-\frac{x}{y}$, with $x+y=1$
$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	$\begin{bmatrix}0&0\\0&0\end{bmatrix}$	All γ

4.2. Symmetric SPEs

To deal with the general case of mixed strategies, we make use of the relationship between SPEs and *U*-game NEs implied by Lemma 2.1 and Theorem 3.5. The first step is to determine the possible combinations of NEs that can be present in the *U*-games for the four states. In general there is a different U-game associated with each state, but because transition probabilities in Γ_I are independent of the present state, expected future rewards and therefore the components of the U-game are also independent of state (assuming stationary policies). More formally, observe that U can be expressed in a way that does not depend on the current state. From Eq. (11):

$$\begin{aligned} U_k(f_1,f_2,\xi,a_1,a_2) &= r_k(a_1,a_2) \\ &\quad +\gamma\sum_{\xi'}\left[V_k(f_1,f_2,\xi')\cdot p(\xi'|\xi,a_1,a_2)\right] \\ &= r_k(a_1,a_2)+\gamma V_k(f_1,f_2,\overline{a_1a_2}), \end{aligned} \quad (21)$$

where $\overline{a_1a_2}$ represents the state following an outcome of (a_1, a_2). Because of this invariance we drop the state ξ, as well as the policies f_1 and f_2, from the notation and index U solely by the actions. This reduction to a single U-game considerably simplifies the problem of determining SPEs.

As mentioned above, the restriction to symmetric SPEs implies that the two players' policies are transposes of each other in the matrix representation used here. This in turn leads the U-game to be symmetric as well, that is

$$\begin{aligned} &U_1(C,C)=U_2(C,C), \quad U_1(C,D)=U_2(D,C), \\ &U_1(D,C)=U_2(C,D), \quad U_1(D,D)=U_2(D,D). \end{aligned} \quad (22)$$

The next step is to characterize the possible NEs for this U-game. In general, the NEs of a 2×2 matrix game can be determined by computing the best replies for each player as a function of the other player's strategy, as shown in the best-reply graph in Fig. 1A. The graph shows each player's optimal probability of choosing the focal action (e.g., Cooperate) as a function of the action probability of the opponent. The points where the two lines intersect are mutual best reply pairs, or NEs.

In the symmetric case, the qualitative form of the best-reply graph is determined by the following two comparisons:

$$U_1(C,D)\gtreqless U_1(D,D), U_1(C,C)\gtreqless U_1(D,C). \quad (23)$$

Each of these can take truth values of '>', '<', or '='. Using Eq. (22), these comparisons determine corresponding comparisons for U_2. The best-reply graph in Fig. 1A assumes $U_1(C, D)<U_1(D, D)$ and $U_1(C, C)>U_1(D, C)$.

Another way to graphically represent the NEs of a two-player game is through the "shape" of the possible payoffs. Fig. 1B shows the range of expected joint payoffs to the two players under all possible strategies, for a reward matrix consistent with the example in Fig. 1A. The vertices of the payoff shape correspond to the four deterministic outcomes; circled points correspond to NEs. Although the exact form of the payoff shape depends on relations other than just those given above, the relations in Eq. (23) are sufficient to fix those aspects of the reward structure that are relevant here, namely the best-reply graph and the qualitative set of NEs.

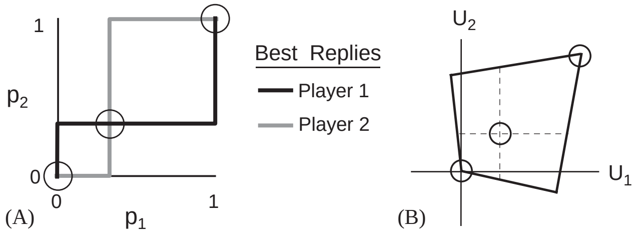


Fig. 1. Example graphical representations of the NEs of a 2×2 game. Circled points in either graph denote NEs. (A) Best-reply graph. Probability of choosing the focal action for player k is denoted by p_k. The best reply curve for player 1 gives the optimal p_1 as a function of p_2 (dark line), and vice versa (light line). (B) Payoff shape. Axes correspond to expected rewards for the two players (here denoted U). Each dashed line indicates the set of outcomes corresponding to the strategy of one player that makes the opponent's payoff independent of his or her action.

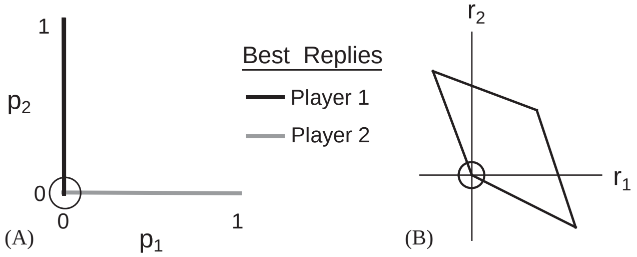


Fig. 2. Best-reply graph (A) and payoff shape (B) for the one-shot Prisoner's Dilemma.

For the one-shot Prisoner's Dilemma, $r_1(C, D)<r_1(D, D)$ and $r_1(C, C)<r_1(D, C)$, and the best-reply graph and payoff shape are as in Figs. 2A and B, respectively. These diagrams both illustrate how the dominance of the defect action leads to (D, D) as the unique NE. However, replacing r-values with U-values, corresponding to a switch from the one-shot to the iterated game, can change the payoff shape in a manner that qualitatively alters the set of NEs. Using the following relation, derived from Eq. (11), we see that each component of U is offset from the corresponding component of r by a scaled-down convex combination of the Us themselves:

$$\begin{aligned} U(a_1,a_2) = r(a_1,a_2)+\gamma\big(&f_1[\overline{a_1a_2}]f_2[\overline{a_1a_2}]U(C,C) \\ &+f_1[\overline{a_1a_2}](1-f_2[\overline{a_1a_2}])U(C,D) \\ &+(1-f_1[\overline{a_1a_2}])f_2[\overline{a_1a_2}]U(D,C) \\ &+(1-f_1[\overline{a_1a_2}])(1-f_2[\overline{a_1a_2}])U(D,D)\big). \end{aligned} \quad (24)$$

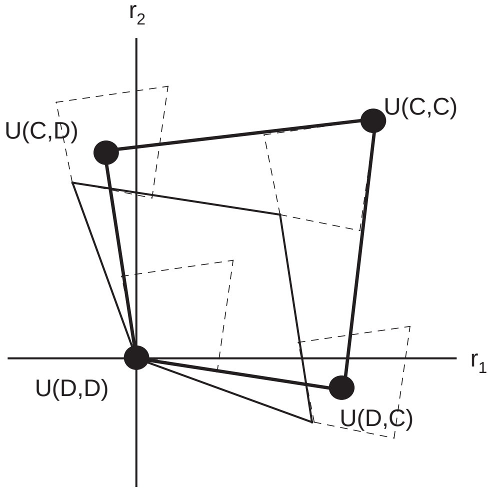


Fig. 3. Example relationship between U (heavy line) and r (medium line) in the IPD. Dashed lines represent shrunken copies of the U-shape. Each small U-shape has its origin aligned with a vertex of the r-shape, and contains the corresponding vertex of the full-scale U-shape (shown by a filled circle).

Fig. 3 illustrates this relationship between r and U. In this diagram, the origin of a shrunken U-graph is aligned with each of the vertices of the r-shape. The space spanned by the vertices of each shrunken U-graph contains the corresponding vertex of the full-scale U-shape. In this example $U(C, C)$ has moved to the right of $U(D, C)$ and above $U(C, D)$, altering the result of the second comparison in Eq. (23) and thus changing the best-reply graph and payoff shape to the forms shown in Fig. 1, which include two new NEs not present in the one-shot PD.

In general, because the manner in which the U-shape deviates from the r-shape is a priori unconstrained, the form of the associated best-reply graph is also unconstrained. Thus there are a total of nine qualitatively different configurations according to the possible results of the relations in Eq. (23), each with a qualitatively different set of NEs, as shown in Table 3. In each of these cases there are multiple potential SPEs, each corresponding to selection of one U-game NE to be played in each of the four states. However, because the values of U are determined by the policies being used, not every such set of assignments will yield the same U-shape as was initially assumed. Thus what remains is to determine which sets are self-consistent, such that the strategy pairs assigned to each state are in fact U-game NEs. This can be done by simultaneously solving the equations giving U as a function of the policies (Eq. (11)) and those giving the policies as functions of U (i.e., those determining the NEs of the U-game, together with the chosen assignments of NEs to strategy pairs), to obtain quantitative values for both. When these calculations are completed (see Appendix B for examples), 13 different SPEs are found, as listed in Table 4. For those SPEs involving mixed strategies, denoted by p, q, r, and s, the values of these action probabilities are functions of the game parameters x, y, and γ; the specific relationships are given in Appendix C.

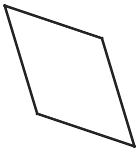
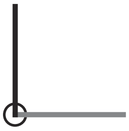
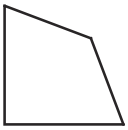
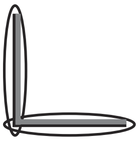
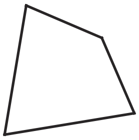
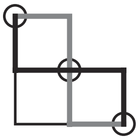
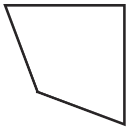
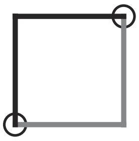
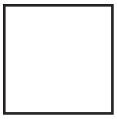
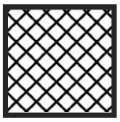
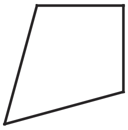
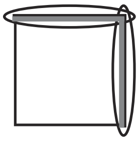
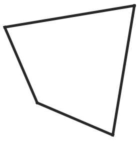
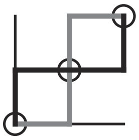
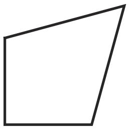
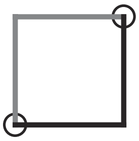
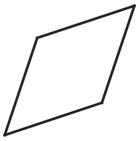
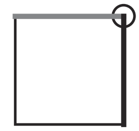
All of the first nine SPEs listed in Table 4 (including the subsequent two duplicates) are repetitions or generalizations of deterministic SPEs #1 and #2 from Table 2, in that they can be put into the form of one of these earlier two by considering the case of either $p = 0$ or $p = 1$. The next two are similarly extensions of deterministic SPE #3. The equilibrium involving all mixed strategies— $\begin{bmatrix} p & q \\ r & s \end{bmatrix}$ vs. $\begin{bmatrix} p & r \\ q & s \end{bmatrix}$ —is a generalization of deterministic SPE #5, TFT vs. TFT. Just as in the deterministic case, these policies make the opponent's policy irrelevant to his or her expected total reward (see Appendix C for the constraints relating p, q, r, and s). Finally, we have once again the all-defect equilibrium, which is the analogue of the unique NE of the single-trial game.

4.3. Connectedness of all-defect

A further question to ask regarding SPEs in the IPD is whether and when cooperative policies can arise when both players initially use the all-defect policy. Because the iterated game is defined to go on indefinitely, and policies are defined as determining actions at all potential future stages, this question requires a dual timescale perspective, in which players' policies (or the players themselves, under an evolutionary interpretation) change at an infinitesimal rate relative to the progression of the game. In this framework one can also consider the game parameters x, y, and γ to vary along the slow timescale. Thus the question becomes whether the all-defect SPE is connected to any other (necessarily semi-cooperative) SPEs within the 11-dimensional space defined by the two players' policies and the three game parameters.

It can be shown that none of the mixed SPEs derived here degenerates to all-defect under any values of the payoff parameters satisfying the strict requirements $x<0$ and $y>1$ (even in the limit $\gamma \to 1$). However, the topological notion of connectedness requires that we also consider the continuation of the solutions onto the boundary of the parameter space, by allowing $x \to 0, y \to 1$, or both. In the case of $\begin{bmatrix} 0 & 0 \\ p & 0 \end{bmatrix}$ vs. $\begin{bmatrix} 0 & p \\ 0 & 0 \end{bmatrix}$, we have $\lim_{x\to 0}(p)=0$ for any $y\geqslant 1$ and $\gamma \in (0,1)$. Thus, allowing x to approach 0 leads to semi-cooperative SPEs arbitrarily close to all-defect (while still remaining within the proper parameter space). A similar situation arises for $\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} p & q \\ r & s \end{bmatrix}$ vs. $\begin{bmatrix} p & r \\ q & s \end{bmatrix}$, which both continuously approach all-defect as (x, y)

Table 3
Best-reply graphs and NEs for all nine symmetric configurations of U

Defining Relations	Example U-shape	Best Replies	NEs
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) < U_1(D,C)$			(0,0)
$U_1(C,D) = U_1(D,D)$ $U_1(C,C) < U_1(D,C)$			(p,0), (0,p)
$U_1(C,D) > U_1(D,D)$ $U_1(C,C) < U_1(D,C)$			(0,1), (p,p), (1,0)
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) = U_1(D,C)$			(0,0), (1,1)
$U_1(C,D) = U_1(D,D)$ $U_1(C,C) = U_1(D,C)$			(p,q)
$U_1(C,D) > U_1(D,D)$ $U_1(C,C) = U_1(D,C)$			(p,1), (1,p)
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) > U_1(D,C)$			(0,0), (p,p), (1,1)
$U_1(C,D) = U_1(D,D)$ $U_1(C,C) > U_1(D,C)$			(0,0), (1,1)
$U_1(C,D) > U_1(D,D)$ $U_1(C,C) > U_1(D,C)$			(1,1)

Note: Variables p and q represent all values in [0,1].

tend jointly to (0, 1).[6] Therefore these three families of SPEs all provide a potential explanation for the emergence of cooperation, under the dual timescale interpretation.

5. Discussion

Given the potential complexity of behavior that could arise in a repeated or stochastic game, one might suppose that people's cognitive limitations would come into play before they were able to attain a globally optimal outcome, even at the individual level. Therefore it would be reasonable to expect a tradeoff between the payoffs that are achieved and the cost of implementing

[6] The solution $\begin{bmatrix} p & q \\ r & s \end{bmatrix}$ vs. $\begin{bmatrix} p & r \\ q & s \end{bmatrix}$ represents a 2-dimensional set of SPEs for any given values of the game parameters; the limit of this set as $(x, y) \to (0, 1)$ includes the all-defect pair.

Table 4
Symmetric 1-back SPEs for the IPD

U-shape	SPEs	
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) > U_1(D,C)$	$\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$ vs. $\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ vs. $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
	$\begin{bmatrix}p & 0\\0 & 0\end{bmatrix}$ vs. $\begin{bmatrix}p & 0\\0 & 0\end{bmatrix}$	$\begin{bmatrix}p & 0\\0 & p\end{bmatrix}$ vs. $\begin{bmatrix}p & 0\\0 & p\end{bmatrix}$
	$\begin{bmatrix}1 & p\\p & 0\end{bmatrix}$ vs. $\begin{bmatrix}1 & p\\p & 0\end{bmatrix}$	$\begin{bmatrix}1 & p\\p & 1\end{bmatrix}$ vs. $\begin{bmatrix}1 & p\\p & 1\end{bmatrix}$
	$\begin{bmatrix}1 & p\\p & p\end{bmatrix}$ vs. $\begin{bmatrix}1 & p\\p & p\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & p\end{bmatrix}$ vs. $\begin{bmatrix}1 & 0\\0 & p\end{bmatrix}$
	$\begin{bmatrix}p & 0\\0 & 1\end{bmatrix}$ vs. $\begin{bmatrix}p & 0\\0 & 1\end{bmatrix}$	
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) = U_1(D,C)$	$\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$ vs. $\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ vs. $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
$U_1(C,D) = U_1(D,D)$ $U_1(C,C) < U_1(D,C)$	$\begin{bmatrix}0 & 0\\p & 0\end{bmatrix}$ vs. $\begin{bmatrix}0 & p\\0 & 0\end{bmatrix}$	
$U_1(C,D) > U_1(D,D)$ $U_1(C,C) < U_1(D,C)$	$\begin{bmatrix}p & 0\\1 & p\end{bmatrix}$ vs. $\begin{bmatrix}p & 1\\0 & p\end{bmatrix}$	
$U_1(C,D) = U_1(D,D)$ $U_1(C,C) = U_1(D,C)$	$\begin{bmatrix}p & q\\r & s\end{bmatrix}$ vs. $\begin{bmatrix}p & r\\q & s\end{bmatrix}$	
$U_1(C,D) < U_1(D,D)$ $U_1(C,C) < U_1(D,C)$	$\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$ vs. $\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$	

Notes: Equilibria are categorized according to the U-shape they induce (note that two are repeated). Here p, q, r, and s each represent some cooperation probability between 0 and 1. See Appendix C for the equations relating these values to the game parameters x, y, and γ.

the policies necessary to achieve them (Aumann, 1981). The present results indicate that this need not be the case, in that mutual restriction to the use of bounded information can be individually rational, thus relieving both players of the incentive to process more complex policies.

This is certainly not to say that computational complexity, in terms of the amount of information or trial-to-trial memory required to implement a policy, is not a factor in people's decisions. Rather, our results simply suggest that this factor will not force players outside the SPE solution set. This is because, even in a case where players might be motivated to sacrifice optimal payoffs for the sake of simpler policies, best-reply universality implies that neither player will choose a more complex policy than the other; that is, both players will restrict themselves to the same complexity class. Applying SPE universality to this class then implies that the outcome is a true SPE.[7] Thus the universality results suggest a way around the primary complications introduced by issues of cognitive economy, including exactly how players compromise between complexity and payoffs, and whether such tradeoffs might lead to non-equilibrium outcomes.

Cognitive economic considerations of the present results may even salvage the SPE as a useful tool for descriptive theory. Specifically, we conjecture that the majority of real human interactions take place at those SPEs involving the least complex policies. Furthermore, these low-information SPEs will in general be the least numerous, because they correspond to the I-games with the fewest numbers of states. Thus the universality theorems turn the previously weak concept of SPE into one with real predictive power.

With regard to analysis of particular games, the universality theorems provide a powerful tool for evaluating the best-reply and SPE criteria, by reducing both optimality requirements to ones that must hold with respect to a much smaller set of deviations (i.e., those following any value of I, rather than following any history). Further, the method shown here allows for the discovery of novel SPEs, often composed of simple and interpretable policies, through various (perhaps independently motivated) choices of the function I. An example is our investigation of mutually optimal 1-back policies in the IPD. The set of SPEs derived includes some well-known policies, such as Grim and TFT, and a number of novel ones that manifest various patterns of semi-cooperative behaviors.

Some particularly interesting results arose in this investigation regarding TFT. This policy was found to be effective in that best replies to it tend to be highly cooperative, yet it is rarely a best reply itself (and then only in cases where player 1's policy is largely irrelevant; see Table 1). The reason for this seeming contradiction is that the power of TFT is restricted to situations in which the opponent is able to assess the contingency between his or her actions and TFT's subsequent responses. This is the case when we consider the best replies to TFT, because the concept of best reply implicitly assumes that the opponent knows the player's policy. However, in the case of the focal player determining the best reply to a fixed policy of the opponent, the situation is quite different. Here, TFT is only effective if the opponent's policy has some intrinsic ability to learn the future effects of its own actions.

Of course, any process by which one player identifies the other's policy during play and learns to respond optimally can be formalized as a policy within the game. In particular, a learning algorithm based on trial-by-trial updating of parameters (e.g., Q-values or association weights) is equivalent to a highly history-dependent policy, for which actions typically depend on all past outcomes via the memory contained in the learned parameters (this is the case even though the updating procedure is based only on recent information). A consequence of the best-reply universality theorem is

[7] A complexity class is given by $\cup_{I \in \mathfrak{I}_n} F_I$, where $\mathfrak{I}_n$ is the set of sufficient and deterministic I that take on no more than n values for a given n. In general, it can be easily shown that any class that is a union of best-reply universal classes is itself best-reply and SPE universal.

that if the opponent's policy is temporally bounded, then the unbounded memory implied by most learning algorithms is unnecessary, as there will always exist a best reply that is itself temporally bounded. This fact highlights a shortcoming of the equilibrium-based approach to rationality in iterated games, in that it denies the need for online learning by essentially assuming that players have foreknowledge of each other's policy. Under this assumption, powerful algorithms for learning and adapting to an unknown policy of the opponent will fare worse than a simple program that happens to "know" in advance what it is up against. Therefore the SPE approach to rationality potentially ignores some of the more interesting dynamics that could arise as a consequence of online learning.[8]

5.1. Universality and machine games

The concepts of restricted SPEs and informationally bounded policies discussed here have close connections to previous investigations of behavioral complexity in repeated games. In particular, the I-game as described here is very similar to the machine game that has been studied by Rubinstein and colleagues (Rubinstein, 1986; Abreu & Rubinstein, 1988; Osborne & Rubinstein, 1994, Chap. 9). A machine game is a repeated game in which each player chooses actions based on the state of some automaton whose updating process depends on the outcomes of previous trials. Typically, the automaton's transition function is defined using only the action of the opponent, ignoring the focal player's action (exceptions can be found in Kalai & Stanford, 1988; Osborne & Rubinstein, 1994). A policy based on the automaton is a function specifying an action (typically *not* a mixed strategy) in the one-shot game for each state of the machine.

It is easy to see how any function I satisfying the determinism axiom is equivalent to an automaton, and therefore the policies and SPEs falling under the approach presented here are precisely those that can arise in machine games.[9] In the case of pure policies this implies a great deal of scope for the present framework, as for any deterministic SPE in a repeated game there exists a machine-based deterministic SPE that produces the same sequence of outcomes (Osborne & Rubinstein, 1994, p. 154; see also Abreu & Rubinstein, 1988).

The most relevant result from the theory of machine games is that of Abreu and Rubinstein (1988), who prove that if one player uses a pure policy based on a finite-state automaton, then the opponent has a pure best reply that is also based on an automaton, and furthermore this automaton can be made to have no more states than that of the first player. This is equivalent to the statement that the class $\bar{M}_n$ of pure policies implementable using an automaton of n or fewer states is best-reply universal. The result can also be seen as a corollary of our best-reply universality theorem, which implies that given a machine for player 2's policy, player 1 has a best reply based on the same machine. In fact, Abreu and Rubinstein's (1988) proof follows essentially these lines, in that they construct a best-reply automaton for player 1 that mimics the dynamics of player 2's machine.

The present results can therefore be viewed as an extension of Abreu & Rubinstein's (1988) theorem, generalizing it in a number of ways. First, the present approach allows mixed policies, and thus implies that M_n (the class of mixed policies based on automata with n or fewer states) is also universal. Because pure SPEs do not necessarily exist at all in an arbitrary repeated game (consider the repetition of a 0-sum game in which there is no NE in pure strategies), the restriction to pure policies can be quite a severe one.

One interesting set of results relating to the issue of pure versus mixed policies concerns the situation in which players use pure policies, but choose them stochastically. Under this scenario, in the case of 0-sum games, Ben-Porath (1993) proves that if one player's policy is sufficiently more complex than the other's, as measured by the number of states required for implementation with an automaton, then the first player can achieve a better outcome than (s)he would obtain if both were limited to the same complexity. (Lehrer, 1988, proves a similar result based on the length of memory of players' policies.) Therefore the universality theorems do not apply to this type of game. Interestingly, the proof uses the construction of an automaton that first discerns the policy being used by the opponent and then selects the best reply to it. Thus this model may be useful in addressing the shortcomings mentioned above regarding modeling of learning processes.

A second contribution of the present approach is that it allows players' strategies to depend on their own past actions. In the case of pure policies this actually makes no difference for the machine model, as the player's past actions are fully determined by the past states of the automaton. Thus any reliance on one's own past actions can be absorbed into the transition function, specifically the dependence of each state on the previous state (this

[8] The difficulty still arises even when both players use learning algorithms. Because such an algorithm updates its parameters trial by trial, the information contained in these parameters satisfies the determinism axiom, placing the class of policies that can be founded on the algorithm within the scope of the universality results. Best-reply universality thus implies that a player cannot gain an advantage by using a more complex learning algorithm than that used by the opponent. However, this analysis again assumes that the player has foreknowledge of the opponent's algorithm. In the absence of this assumption more complex algorithms could well fare better.

[9] Technically, all SPEs considered here must be implemented with all players using the *same* automaton (i.e., same information I), whereas this restriction is not present in machine games. However, given any two policies based on different automata, one can construct the product automaton and base both policies on that machine.

is the approach used in the aforementioned proof of Abreu & Rubinstein, 1988). However, in the case of mixed policies, the outcomes of one's own stochastically determined past actions can provide useful additional information (note that using stochastic machine-state transitions is not sufficient, as they will be uncorrelated with the randomness in the actions). All of the SPEs that arose in our IPD analysis that contain mixed strategies serve as examples of how dependence upon one's own previous actions can allow for optimal behaviors not otherwise achievable.

A third manner in which the present results extend previous work is the application to stochastic games, and the provision of a condition under which the universality properties hold, namely the sufficiency axiom. This axiom qualifies the universality results by stating that information can only be safely ignored if it is ignored by one's opponents *and* it is irrelevant to the current state of the game. This latter condition is trivial in a repeated game, but in a stochastic game it is critical for ensuring that the best-reply and SPE criteria are met.[10] The sufficiency axiom also plays an important role in finite-length games. For example, in the finitely repeated Prisoner's Dilemma the only SPE is all-defect, but if players are restricted to the complexity class M_n for n sufficiently small relative to the length of the game then there exist stable outcomes (restricted SPEs) involving cooperation (Neyman, 1985). Thus M_n is not SPE universal for this game. This discrepancy with the infinite game can be understood in terms of the sufficiency axiom as follows: Because the game has a definite ending point, all trials must be treated as different states in order for the game to be Markov. Under this interpretation the SPE universality theorem would hold, but the sufficiency axiom implies that the machine must have at least as many states as there are steps in the game. Therefore limiting the complexity to a value less than the length of the game prevents the assumptions of the universality theorems from being met. Thus the present framework provides a natural explanation for the critical role of complexity limitations in finite-length games.

Acknowledgments

This work was supported in part by a Pre-doctoral Fellowship to M.J. from the Horace H. Rackham Graduate School at the University of Michigan. The authors thank David Meyer, Scott Page, and Thad Polk for helpful comments on an earlier draft of this paper.

Appendix A. Calculation of the best reply to a deterministic policy

The following is a representative example of the calculation of player 1's optimal pure 1-back reply to $f_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, as a function of the game parameters x, y, and γ. Optimality is defined in terms of the requirement given in Definition 2.1, which for stationary policies is equivalent to simultaneous maximization of all four state values $V_1(\xi)$.

First consider what will happen if player 1 uses the all-defect strategy: If the game starts in state CC or DD, player 2 will cooperate on the first trial while player 1 defects, giving player 1 a reward of y and putting the players in state DC. Both players will now defect, yielding a reward of 0 and putting the game into state DD. The game will then continue to alternate between DC and DD, implying

$$V_1(CC) = V_1(DD) = \frac{y}{1-\gamma^2}. \tag{A.1}$$

Similar direct calculations yield

$$V_1(CD) = V_1(DC) = \frac{\gamma y}{1-\gamma^2}. \tag{A.2}$$

These values can now be used as lower bounds on the V-values for the optimal policy. Next consider player 1's action in state CD: If (s)he were to cooperate, (s)he would receive a reward of x (as the opponent would be defecting), and the game would remain in CD indefinitely. Thus we would have

$$V_1(CD) = \frac{x}{1-\gamma} < \frac{\gamma y}{1-\gamma^2} \tag{A.3}$$

which would be inferior to all-defect. Therefore the optimal policy for the present case has $f_l[CD]=0$. Now consider $f_l[DC]$: If $f_l[DC]=1$, then DC would always lead to CD followed by DD, whereas if $f_l[DC]=0$ then DC would lead directly to DD. Thus

$$\begin{aligned} f_1[DC] = 1 &\Rightarrow V_1(DC) = x + \gamma^2 V_1(DD), \\ f_1[DC] = 0 &\Rightarrow V_1(DC) = \gamma V_1(DD). \end{aligned} \tag{A.4}$$

By Eq. (A.1), the optimal policy satisfies $V_1(DD)>0$, and this along with $x<0$ and $0\leqslant\gamma<1$ implies that the second expression for $V_1(DC)$ in Eq. (A.4) is strictly greater than the first. Therefore $f_1[DC]=0$. This also implies $V_1(DC)=\gamma V_1(DD)$, so maximization of $V_1(DD)$ will ensure that of $V_1(DC)$. It can be similarly shown that maximizing $V_1(DD)$ also leads to maximization of $V_1(CD)$.

[10] Consider the case where I is a constant function: I is deterministic but not sufficient, and any policy based on I will dictate the same strategy regardless of the current state. If the opponent uses such a policy, and the one-shot best reply to this strategy differs across states (because of differing reward structures), then knowledge of the current state would allow for increased payoffs. Thus F_I is not universal.

With only two components of the optimal policy left undetermined (and thus only four possible policies remaining), and only two V-values to consider, we now proceed with direct calculations:

$$\begin{aligned}
&f_1[CC]=0, f_1[DD]=0 \Rightarrow V_1(CC)=\frac{y}{1-\gamma^2}, \quad V_1(DD)=\frac{y}{1-\gamma^2},\\
&f_1[CC]=0, f_1[DD]=1 \Rightarrow V_1(CC)=\frac{y+\gamma^2}{1-\gamma^3}, \quad V_1(DD)=\frac{1+\gamma y}{1-\gamma^3},\\
&f_1[CC]=1, f_1[DD]=0 \Rightarrow V_1(CC)=\frac{1}{1-\gamma}, \quad V_1(DD)=\frac{y}{1-\gamma^2},\\
&f_1[CC]=1, f_1[DD]=1 \Rightarrow V_1(CC)=\frac{1}{1-\gamma}, \quad V_1(DD)=\frac{1}{1-\gamma}.
\end{aligned} \tag{A.5}$$

When $\gamma < y-1$, the potential state values in Eq. (A.5) are ordered as follows:

$$\frac{y}{1-\gamma^2} > \frac{y+\gamma^2}{1-\gamma^3} > \frac{1+\gamma y}{1-\gamma^3} > \frac{1}{1-\gamma}. \tag{A.6}$$

If $\gamma > y-1$ then these inequalities are reversed. Therefore if $\gamma < y-1$ then the optimal policy is $\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$, and if $\gamma > y-1$ (which is only possible if $y<2$) then the best reply is $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$. Finally, if $\gamma = y-1$ then all four policies considered above— $\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$, $\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}$, $\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$ and $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$—lead to the same set of expected payoffs, and thus all four are optimal.

The dynamics behind the transition from $\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$ to $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ can also be understood through a consideration of the U-values. Calculation of U_1 for the above four policies yields:

$$\begin{aligned}
f=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix} &\Rightarrow U_1=\begin{bmatrix}1+\frac{\gamma y}{1-\gamma^2} & x+\frac{\gamma^2 y}{1-\gamma^2}\\ \frac{y}{1-\gamma^2} & \frac{\gamma y}{1-\gamma^2}\end{bmatrix},\\
f=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} &\Rightarrow U_1=\begin{bmatrix}\frac{1+\gamma y}{1-\gamma^3} & x+\frac{\gamma^2+\gamma^3 y}{1-\gamma^3}\\ \frac{y+\gamma^2}{1-\gamma^3} & \frac{\gamma+\gamma^2 y}{1-\gamma^3}\end{bmatrix},\\
f=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} &\Rightarrow U_1=\begin{bmatrix}\frac{1}{1-\gamma} & x+\frac{\gamma^2 y}{1-\gamma^2}\\ \frac{y}{1-\gamma^2} & \frac{\gamma y}{1-\gamma^2}\end{bmatrix},\\
f=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} &\Rightarrow U_1=\begin{bmatrix}\frac{1}{1-\gamma} & x+\frac{\gamma^2}{1-\gamma}\\ y+\frac{\gamma^2}{1-\gamma} & \frac{\gamma}{1-\gamma}\end{bmatrix}.
\end{aligned} \tag{A.7}$$

For all of these policies we have $U_1(D, D) > U_1(C, D)$, corresponding to the fact that the best reply to an upcoming defection by player 2 is to defect as well, regardless of x, y, or γ (this is consistent with $f_1[CD]=f_1[DC]=0$, since $f_2[CD]=f_2[DC]=0$). Furthermore, all four cases satisfy $U_1(C, C) > U_1(D, C)$ if and only if $\gamma > y-1$. This is consistent with the fact that the best reply matches cooperation with cooperation ($f_1[CC]=f_1[DD]=1$) if $\gamma > y-1$, and counters cooperation with defection ($f_1[CC]=f_1[DD]=0$) if $\gamma < y-1$. When $\gamma = y-1$, and $U_1(C, C) = U_1(D, C)$ for the policies under consideration, the response to the opponent's cooperation is irrelevant, which is why all four policies are optimal.

Appendix B. Determination of symmetric mixed SPEs

The following is a sketch of the calculations involved in determining the set of symmetric SPEs associated with the U-configuration defined by $U_1(C, C) > U_1(D, C)$ and $U_1(D, D) > U_1(C, D)$. The NEs associated with this type of payoff matrix are (0, 0), (1, 1), and (p, p) for a unique $p \in (0, 1)$, given by

$$p=\frac{U_1(D,D)-U_1(C,D)}{U_1(C,C)-U_1(D,C)-U_1(C,D)+U_1(D,D)}. \tag{B.1}$$

Therefore we need to search through all policy pairs (f, f^{T}) satisfying $(f[\xi], f^{\mathrm{T}}[\xi]) = (0, 0)$, (1, 1), or ($p$, p) for every state ξ. Note that the symmetry present in all three NEs, along with the symmetry assumed between the players' policies, implies $f[CD]=f^{\mathrm{T}}[CD]=f[DC]$. Thus there are three assignments to be made, each with three choices, yielding 27 policy pairs that are potential SPEs.

The relations assumed among the components of U also have implications for V, via the following simplification of Eq. (11):

$$U_1(a_1, a_2) = r_1(a_1, a_2) + \gamma V_1(\overline{a_1 a_2}). \tag{B.2}$$

In particular, given that $r_1(C, C) < r_1(D, C)$, the assumption $U_1(C, C) > U_1(D, C)$ implies

$$V_1(CC) > V_1(DC). \tag{B.3}$$

Furthermore, the symmetry assumed for U and for the policies implies both $V_1(\xi) = V_2(\xi)$ for every ξ and $V_k(CD) = V_k(DC)$ (see Eq. (12)). Therefore for the remainder of this appendix the subscript on V will be dropped and $V(CD)$ and $V(DC)$ will be used interchangeably.

The number of policy pairs under consideration can be reduced by an initial analysis of the state values associated with the three NE strategy pairs, in the following manner. For each U-game NE (b, b), with $b = 0, 1$, or p, define the *equilibrium value* $W^{(b,b)}$ by

$$W^{(b,b)} = E\left[\sum_{t \geqslant \tau} \gamma^{t-\tau} r_1^t \,|\, P[a_1^\tau = C] = P[a_2^\tau = C] = b\right]. \tag{B.4}$$

Note that $W^{(b,b)}$ is equal to the state value $V(\xi)$ for any ξ for which $(f[\xi], f^{\mathrm{T}}[\xi]) = (b, b)$ (see Eq. (10)). Therefore constraints on the correspondences between state values V and equilibrium values W can provide constraints on the policy f (e.g., $V(\xi) \neq W^{(b,b)}$ implies $f[\xi] \neq b$).

From Eq. (12) and the relationship between V and W, $W^{(b,b)}$ is equal to the expected payoff to player 1 (and to player 2) for the one-shot U-game NE corresponding to the strategy pair (b,b):

$$\begin{aligned} W^{(1,1)} &= U_1(C,C), \\ W^{(0,0)} &= U_1(D,D), \\ W^{(p,p)} &= p^2 U_1(C,C) + p(1-p)U_1(C,D) \\ &\quad + p(1-p)U_1(D,C) + (1-p)^2 U_1(D,D). \end{aligned} \tag{B.5}$$

Because (p,p) is a mixed-strategy NE, player 1 gets the same expected payoff regardless of his or her strategy, and therefore in particular the expected NE payoff is equal to the expected payoff under either pure action:

$$\begin{aligned} W^{(p,p)} &= pU_1(C,C) + (1-p)U_1(C,D) \\ &= pU_1(D,C) + (1-p)U_1(D,D). \end{aligned} \tag{B.6}$$

Here player 2 is still assumed to cooperate with probability p, whereas player 1 either cooperates (middle expression) or defects (RHS) with certainty.

Assume for the moment that $W^{(p,p)} \geqslant U_1(C, C)$. By the first equality in Eq. (B.6), this implies $U_1(C, C) \leqslant U_1(C, D)$. Using the initial assumptions $U_1(C, C) > U_1(D, C)$ and $U_1(C, D) < U_1(D, D)$, this implies that the components of U are ordered as $U_1(D, D) > U_1(C, D) \geqslant U_1(C, C) > U_1(D, C)$. The second equality in Eq. (B.6) (LHS = RHS) now yields $W^{(p,p)} < U_1(D, D)$. Therefore, referring back to the hypothetical assumption $W^{(p,p)} \geqslant U_1(C, C)$, we can conclude that either $W^{(p,p)} < U_1(C, C)$ or $W^{(p,p)} < U_1(D, D)$. By Eq. (B.5) this is equivalent to:

$$W^{(p,p)} < W^{(0,0)} \text{ or } W^{(p,p)} < W^{(1,1)}. \tag{B.7}$$

The possible orderings of the Ws consistent with Eq. (B.7) are broken into three cases: $W^{(1,1)} > W^{(p,p)} \geqslant W^{(0,0)}$; $W^{(1,1)} > W^{(0,0)} > W^{(p,p)}$; and $W^{(0,0)} > W^{(p,p)}$, $W^{(0,0)} \geqslant W^{(1,1)}$.

Case 1: $W^{(1,1)} > W^{(p,p)} \geqslant W^{(0,0)}$. In this case, Eq. (B.3) rules out all but three choices for the pair $(V(CC), V(DC))$: $(W^{(1,1)}, W^{(0,0)})$, $(W^{(1,1)}, W^{(p,p)})$, and $(W^{(p,p)}, W^{(0,0)})$. This implies that $(f[CC], f[DC])$ must be equal to (1, 0), (1, p), or (p, 0). Crossing these options with the three choices for $f[DD]$ yields 9 policy pairs that must be considered: $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & p \end{bmatrix}$ vs. $\begin{bmatrix} 1 & 0 \\ 0 & p \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ vs. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & p \\ p & 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 & p \\ p & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$ vs. $\begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$, $\begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$ vs. $\begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$, $\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix}$, and $\begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix}$.

Case 2: $W^{(1,1)} > W^{(0,0)} > W^{(p,p)}$ In this case the only way to satisfy Eq. (B.3) that was not listed in Case 1 is with $(f[CC], f[DC]) = (0, p)$ and hence $(V(CC), V(DC)) = (W^{(0,0)}, W^{(p,p)})$. Furthermore, the assumption $W^{(0,0)} > W^{(p,p)}$, or equivalently $U_1(D, D) > W^{(p,p)}$, implies via Eq. (B.6) that $U_1(D, D) > U_1(D, C)$. This along with $r_1(D, D) < r_1(C, D)$ implies $V(DD) > V(DC)$ by Eq. (B.2). Therefore $V(DD) \neq W^{(p,p)}$, implying $f[DD] \neq p$. This leaves two new possibilities: $\begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix}$ vs. $\begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & p \\ p & 1 \end{bmatrix}$ vs. $\begin{bmatrix} 0 & p \\ p & 1 \end{bmatrix}$.

Case 3: $W^{(0,0)} > W^{(p,p)}$, $W^{(0,0)} \geqslant W^{(1,1)}$. The assumption $W^{(0,0)} \geqslant W^{(1,1)}$ is equivalent to $U(D, D) \geqslant U(C, C)$ by Eq. (B.5). Using $r_1(D, D) < r_1(C, C)$, Eq. (B.2) implies $V(DD) > V(CC)$. Taking into account Eq. (B.3) gives the full ordering $V(DD) > V(CC) > V(DC)$. The only two possibilities for $((V(DD), V(CC), V(DC))$ are thus $(W^{(0,0)}, W^{(1,1)}, W^{(p,p)})$ and $(W^{(0,0)}, W^{(p,p)}, W^{(1,1)})$, implying that $(f[DD], f[CC], f[DC]) = (0, 1, p)$ or $(0, p, 1)$. The first of these was listed in Case 1, so the final candidate is $\begin{bmatrix} p & 1 \\ 1 & 0 \end{bmatrix}$ vs. $\begin{bmatrix} p & 1 \\ 1 & 0 \end{bmatrix}$.

The final step is to check directly which of the reduced set of 12 candidates are true SPEs, by determining whether the U-game generated by each policy pair has all four constituent strategy pairs as NEs. Two examples are shown here.

Example A:

$$\begin{bmatrix} 1 & p \\ p & p \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$$

With this policy pair mutual cooperation perpetuates itself indefinitely, so we see readily that

$$U_1(C,C) = \frac{1}{1-\gamma}. \tag{B.8}$$

Using Eqs. (B.2) and (B.8) to substitute for the components of U in Eq. (B.6), and then replacing $V(CD)$, $V(DC)$, and $V(DD)$ with $W^{(p,p)}$, yields the following condition for (p, p) to be a U-game NE:

$$\begin{aligned} W^{(p,p)} &= \frac{p}{1-\gamma} + (1-p)(x + \gamma W^{(p,p)}) \\ &= p(y + \gamma W^{(p,p)}) + (1-p)\gamma W^{(p,p)} \end{aligned} \tag{B.9}$$

$$\Rightarrow W^{(p,p)} = \frac{p + (1-p)(1-\gamma)x}{(1-\gamma)(1-\gamma+\gamma p)} \text{ and } W^{(p,p)} = \frac{py}{(1-\gamma)} \tag{B.10}$$

$$\Rightarrow \gamma y p^2 + ((1-\gamma)(x+y)-1)p - (1-\gamma)x = 0. \tag{B.11}$$

Under the restriction to $0 \leqslant p \leqslant 1$ and $0 \leqslant \gamma < 1$, there is one solution curve, given by

$$p = \frac{1-(1-\gamma)(x+y) \pm \sqrt{(1-\gamma)^2(x^2+y^2)+2(1-\gamma^2)xy-2(1-\gamma)(x+y)+1}}{2\gamma y}, \tag{B.12}$$

with boundary points at $(\gamma, p) = (1, 0)$ and $(1, 1/y)$. This curve represents the set of pairs (γ, p) for which (p, p) is a U-game NE.

The other thing that needs to be checked is that the (1, 1) strategy pair associated with state CC is also an NE, which is equivalent to $U_1(C, C) \geqslant U_1(D, C)$. On the solution curve given by Eq. (B.12), where (p, p) is an NE, this is equivalent to the requirement $U_1(C, D) \leqslant U_1(D, D)$ by Eq. (B.6). Using Eq. (B.2), we have

$$U_1(C,D) = x + \gamma V(CD) = x + \gamma W^{(p,p)} \tag{B.13}$$

and

$$U_1(D,D) = 0 + \gamma V(DD) = \gamma W^{(p,p)}. \tag{B.14}$$

Since $x<0$, this implies $U_1(C, D) < U_1(D, D)$, and thus (1, 1) is an NE everywhere on the curve of Eq. (B.12). Therefore $\begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$vs.$\begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$ is an SPE for all γ such that the values of p given by Eq. (B.12) are real and at least one lies in [0, 1]. It can be shown algebraically that this condition is satisfied whenever $\gamma \geqslant \gamma^*$, with the critical value $\gamma^* \in (0,1)$ given by

$$\gamma^* = \frac{x(x-1)+y(y-1)+2\sqrt{xy(x-1)(y-1)}}{(x-y)^2} \tag{B.15}$$

Example B:

$$\begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix} \text{vs.} \begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix}.$$

In order for (0, 0) to be an NE for the U-game, we must have $U_1(C, D) \leqslant U_1(D, D)$. This along with Eq. (B.6) (the criterion for (p, p) to be an NE) implies $U_1(D, C) \leqslant U_1(C, C)$. Since state CC is followed by eternal defection, $U_1(C, C) = 1$. Therefore by Eq. (B.2):

$$V(DC) = \frac{1}{\gamma}(U(D,C)-y) \leqslant \frac{1}{\gamma}(U(C,C)-y) = \frac{1}{\gamma}(1-y) < 0. \tag{B.16}$$

However, by choosing $f_1[\xi] \equiv 0$ player 1 can always guarantee every state value to be nonnegative (for any policy of player 2). This contradiction negates the assumption that (0, 0) and (p, p) are both U-game NEs, and implies that there is no SPE of this type.

Appendix C. Characterization of all symmetric SPEs

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Exists $\forall x,y$. Requires $\gamma \geqslant 1 - 1/y$.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Exists iff $y<2$. Requires $\gamma \geqslant \gamma - 1$.

$$\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix} \text{vs.} \begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$$

$$p = \frac{x+y-1+\sqrt{x^2+y^2+2(1-2\gamma)xy-2x-2y+1}}{2\gamma y}.$$

Exists $\forall x,y$. Requires $\gamma \geqslant 1 - 1/y$.

$$\begin{bmatrix} 1 & p \\ p & 0 \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & p \\ p & 0 \end{bmatrix}$$

$$\begin{aligned} p = {} & [2\gamma(1-(1-\gamma)x+(1-\gamma)y)]^{-1} \\ & \cdot \Big[1-(1-\gamma^2)x-(1-\gamma)^2 y \\ & \pm\sqrt{(1-\gamma)^4(x^2+y^2)+2(1-\gamma)^2(1+2\gamma-\gamma^2)xy-2(1-\gamma)^2(x+y)+1}\Big] \end{aligned}$$

and $p \leqslant \frac{-x}{y-x-1}$.
Exists $\forall x,y$.

$$\begin{bmatrix} 1 & p \\ p & p \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & p \\ p & p \end{bmatrix}$$

$$p = \frac{1-(1-\gamma)(x+y) \pm \sqrt{(1-\gamma)^2(x^2+y^2)+2(1-\gamma^2)xy-2(1-\gamma)(x+y)+1}}{2\gamma y}.$$

Exists $\forall x, y$. Requires $\gamma \geqslant \frac{x(x-1)+y(y-1)+2\sqrt{xy(x-1)(y-1)}}{(x-y)^2}$.

$$\begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$$

$$\begin{aligned} p = {} & [\gamma(y-x+1)]^{-1} \cdot \Big[1+2\gamma-(1+\gamma)x-(1-\gamma)y \\ & \pm\sqrt{(1-\gamma)^2(x^2+y^2)+2(1+2\gamma-\gamma^2)xy-2(1+\gamma)(x+y)+1+4\gamma}\Big]. \end{aligned}$$

Exists iff $y \leqslant 1 + 1/(4(1-x))$.

$$\begin{bmatrix} 1 & 0 \\ 0 & p \end{bmatrix} \text{vs.} \begin{bmatrix} 1 & 0 \\ 0 & p \end{bmatrix}$$

$$\begin{aligned} p = {} & [2\gamma(1-(1-\gamma)x-\gamma y)]^{-1} \cdot \Big[(1-\gamma)^2 x+(1-\gamma^2)y-1 \\ & -\sqrt{(1-\gamma^2)^2(x^2+y^2)+2(1-\gamma)(1-\gamma+\gamma^2+\gamma^3)xy-2(1-\gamma^2)(x+y)+1}\Big]. \end{aligned}$$

Exists $\forall x,y$.

$$\begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix} \text{vs.} \begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix}$$

$$p = [2\gamma(1 - x + y)]^{-1} \cdot \Big[(1-\gamma)x + (1+\gamma)y - 1 + \sqrt{(1+\gamma)^2(x^2+y^2) + 2(1-2\gamma-\gamma^2)xy - 2(1+\gamma)(x+y) + 1}\Big].$$

Exists iff $y < 2$. Requires $\gamma \geqslant y - 1$.

$$\begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix} \text{ vs. } \begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix}$$

$$p = [2\gamma(-\gamma x + (1+\gamma)y)]^{-1} \cdot ((1+\gamma-\gamma^2)x + (1+\gamma+\gamma^2)y - 1 - 2\gamma + [(1+\gamma+\gamma^2)^2(x^2+y^2) + 2(1-5\gamma^2-2\gamma^3-\gamma^4)xy - 2(1+3\gamma+\gamma^2)(x+y) + (1+2\gamma)^2]^{1/2}).$$

Exists iff $y < 2$. Requires $\gamma \geqslant y - 1$.

$$\begin{bmatrix} 0 & 0 \\ p & 0 \end{bmatrix} \text{ vs. } \begin{bmatrix} 0 & p \\ 0 & 0 \end{bmatrix}$$

$$p = -\frac{x}{\gamma y}.$$

Exists iff $x + y > 0$. Requires $\gamma \geqslant -x/y$.

$$\begin{bmatrix} p & 0 \\ 1 & p \end{bmatrix} \text{ vs. } \begin{bmatrix} p & 1 \\ 0 & p \end{bmatrix}$$

$$p = [2\gamma(1 + \gamma - x + y)]^{-1} \Big[x + (1+2\gamma)y - 1 + \gamma^2 - \sqrt{(1+4\gamma)(x^2+y^2) + 2(1+2\gamma^2)xy - 2(1+\gamma)^2(x+y) + (1-\gamma^2)^2}\Big].$$

Exists iff $x + y > 0$. Requires $\gamma \geqslant -x/y$.

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix} \text{ vs. } \begin{bmatrix} p & r \\ q & s \end{bmatrix}$$

$$p = \frac{(1+\gamma s)y + \gamma q - \gamma s - 1}{\gamma y}, \quad r = \frac{(\gamma q - \gamma s - 1)x + \gamma s y}{\gamma y}.$$

Requires

$$\frac{(\gamma s + 1)x + \gamma(1-s)y}{\gamma x} \leqslant q \leqslant \frac{(-\gamma s - 1 + \gamma)y + 1 + \gamma s}{\gamma},$$

which requires $\gamma \geqslant \max\{1 - 1/y,\ 1 - 1/(1-x)\}$. Exist solutions $\forall x, y$.

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ vs. } \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Exists $\forall x, y, \gamma$.

References

Abreu, D., & Rubinstein, A. (1988). The structure of Nash equilibrium in repeated games with finite automata. *Econometrica*, *56*(6), 1259–1281.

Aumann, R. J. (1981). Survey of repeated games. In: V. Bohm and H. Nachtkamp (Eds.), *Essays in game theory and mathematical economics in honor of Oskar Morgenstern*. Zurich: Bibliographisches Institut.

Aumann, R. J., & Shapley, L. S. (1994). Long-term competition: A game-theoretic analysis. In Megiddo, N. (Ed.), *Essays in game theory*. New York: Springer.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Ben-porath, E. (1993). Repeated games with finite automata. *Journal of Economic Theory*, *59*(1), 17–32.

Blackwell, D. (1965). Discounted dynamic programming. *The Annals of Mathematical Statistics*, *36*(1), 226–235.

Filar, J. A., & Vrieze, K. (1997). *Competitive Markov decision processes*. New York: Springer.

Fink, A. M. (1964). Equilibrium in a stochastic *N*-person game. *Journal of Science in Hiroshima University, Series A-I*, *28*, 89–93.

Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *Review of Economic Studies*, *38*, 1–12.

Fudenberg, D., & Maskin, E. S. (1986). The folk theorem in repeated games with discounting or incomplete information. *Econometrica*, *54*, 533–554.

Fudenberg, D., & Maskin, E. S. (1991). On the dispensability of public randomization in discounted repeated games. *Journal of Economic Theory*, *53*, 428–438.

Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. In Shavlik, J. (Ed.), *Proceedings of the 15th international conference on machine learning* (pp. 242–250). San Francisco, CA: Morgan Kaufmann.

Jones, M. (2003). *Temporal information and adaptive nationality*. Ph.D. thesis, University of Michigan. [Abs., *Dissertation Abstracts International* 64 (2-B), 979].

Kalai, E., & Stanford, W. (1988). Finite rationality and interpersonal complexity in repeated games. *Econometrica*, *56*(2), 397–410.

Lehrer, E. (1988). Repeated games with stationary bounded recall strategies. *Journal of Economic Theory*, *46*(1), 130–144.

Luce, D., & Raiffa, H. (1957). *Games and decisions*. New York: Dover Publications (pp. 94–102).

Milinski, M. (1987). Tit-for-tat in sticklebacks and the evolution of cooperation. *Nature*, *325*(6103), 433–435.

Nash, J. F. (1950). Equilibrium points in *N*-person games. *Proceedings of the National Academy of Sciences*, *36*, 48–49.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Neyman, A. (1985). Bounded complexity justifies cooperation in the finitely repeated Prisoners Dilemma. *Economics Letters*, *19*(3), 227–229.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, MA: MIT Press.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's Dilemma: A study in conflict and cooperation*. Ann Arbor, MI: The University of Michigan Press.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory*, *21*, 1–9.

Rubinstein, A. (1986). Finite automata play repeated prisoner's dilemma. *Journal of Economic Theory*, *39*(1), 83–96.

Rubinstein, A. (1994). Equilibrium in supergames. In Megiddo, N. (Ed.), *Essays in game theory*. New York: Springer.

Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft*, 121, 301–324 & 667–689.

Stahl, D. O. (1991). The graph of Prisoner's Dilemma supergame payoffs as a function of the discount factor. *Games and Economic Behavior*, *3*, 368–384.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. Ph.D. thesis, Cambridge University.

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature*, *308*(5955), 181–184.