Rationality and Bounded Information in Repeated Games

matt jones and Jun Zhang
Department of Psychology, University of Michigan

Address for correspondence:
matt jones
Department of Psychology
The University of Texas at Austin
1 University Station A8000
Austin, Texas 78712
Email: mattj@psy.utexas.edu
Tel: (512)472-8340
Fax: (512)471-5935
Abstract

Actions in a repeated or stochastic 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 with respect to some limited information about the game’s history will coincide with the universally optimal policy, provided that the information used by all players satisfies two widely inclusive axioms. It is then shown how this result allows explicit calculation of subgame-perfect equilibria (SPEs) for any stochastic game. The technique is applied to the iterated Prisoner’s Dilemma for the case of 1-back memory, and two sets of SPEs are derived. These solutions exhibit varying degrees of (individually rational) cooperation, which can be seen via the present framework to be a direct result of repeated interaction.

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 size and character of the set of rational outcomes in a game theoretic scenario. For instance in the Prisoner’s Dilemma, in which the Defect strategy 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 the analogous concept – subgame-perfect equilibrium (SPE) – in iterated games. There are a number of strong results concerning the expected total rewards that can result from the SPEs of a repeated game (Rubinstein, 1979, 1994; Fudenberg & Maskin, 1986; Aumann & Shapley, 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. Clearly, playing an NE of the one-shot game on each trial constitutes an SPE of the repeated game, but the set of SPEs encompasses much more than just these stationary policies (see, e.g., Osborne & Rubinstein, 1994).

The state of knowledge regarding SPEs is even sparser in the domain of stochastic games, a generalization of repeated games in which 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. In a stochastic game, play of a one-shot NE in each state is by no means guaranteed to produce an SPE, as the future rewards associated with actions’ effects on subsequent trials can offset the equilibrium in immediate payoffs. If players
are assumed to use stationary policies, that is their action probabilities depend only on the state and not on past outcomes, then 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) show that a profile of stationary policies for all players 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 strategies in any stochastic game (Fink, 1964), very little is known about the structure of these stationary SPEs (Filar & Vrieze, 1997, p. 230).

Beyond the question of SPEs, there is the more general issue of determining the best reply to the policies of one’s opponent(s). Given the uncountable class of policies in a repeated or stochastic game, it may appear that determination of the universally optimal solution is often extremely difficult. Thus human players may be 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 the total number of players who performed a particular action (such as Cooperate) on each trial, without regard for who did what. Importantly, 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; 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 that base actions solely on $I$, the player has a 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 (or rational) behavior. Consequently the restriction to finite memory and other types of bounded rationality are seen to be often unboundedly rational.

The main proof presented in this article is based on constructing an expanded stochastic game, based on the original (repeated or stochastic) game, in which the states of the original game are split into substates according to the value of $I$. The construction implies a payoff-preserving isomorphism between policies in the two games, in which stationary policies in the expanded “$I$-game” correspond to history-dependent policies in the original game that base actions solely on $I$. Furthermore, the preservation of payoffs implies that best replies and SPEs in the $I$-game respectively correspond to best replies and SPEs in the original game. The next step in the proof relies on a result of Filar and Vrieze (1997), stating that if all players in a stochastic game use stationary policies then no one has incentive to use a non-stationary policy. This fact, together with the properties of the expanded $I$-game, implies a pair of “universality” results. 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 (which are the stationary policies in the $I$-game) are true SPEs within the full policy space. Therefore if all players choose strategies on each trial using only the reduced information $I$, 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. 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 $n \geq 0$, 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.

Following these universality results, we present an analytical method for deriving all stationary SPEs of a stochastic game (such as the $I$-game just described), based on the qualitative form of the previously defined $U$-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 each state (recall that in a stationary-policy SPE the strategy profile in each state must be a $U$-game NE; Filar & Vrieze, 1997). Once such a set of assignments been made, the policies and $U$-values can be quantitatively determined. If the values of $U$ that arise are consistent with the original qualitative assumptions (so that the strategies in each state do in fact comprise $U$-game NEs), then the policy profile is indeed 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 strategies. This approach will likely be computationally infeasible for more complex games, as the number of cases to be checked grows exponentially with the number of states and actions, but as we show here in a case of four states and two players, each with
two actions in each state, it is reasonable for smaller games. When applied to the
expanded I-game defined above, this method allows derivation of all SPEs in a repeated
or stochastic game whose constituent policies all 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 approach 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).
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 5 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 10
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 $k$ players each with $n_k$ actions. For ease of exposition we restrict to the situation $k = 2$; however, the central theorems 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 $b = \sum a_i a_i \in B_k$, $b(a)$ denotes the probability $\alpha_i$ 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 $E_{b_1, b_2}[r_k(a_1, a_2)] = \sum_{a_1 \in A_1, a_2 \in A_2} b_1(a_1) b_2(a_2) \cdot r_k(a_1, a_2)$ 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:

$$\begin{bmatrix}
C & D \\
(1, 1) & (x, y) \\
(0, 0) & (y, x)
\end{bmatrix}$$

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 Equation 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:

$$\forall b_1 \in B_1: \ r_1(b_1, b_2^*) \leq r_1(b_1^*, b_2^*)$$
$$\forall b_2 \in B_2: \ r_2(b_1^*, b_2) \leq r_2(b_1^*, b_2^*).$$

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

$$H = \{(a_1^0, a_2^0, a_1^1, a_2^1, \ldots, a_1^{t-1}, a_2^{t-1}) \mid t \geq 0, \ \forall k \forall \tau [a_k^\tau \in A_k]\}$$

(where $t = 0$ corresponds to the initial history $h = \emptyset$) then a policy for player $k$ can be written as...
\( f: H \rightarrow B_k. \) \hspace{1cm} (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 which yields an expected total payoff to each player at least as great as that which the player could achieve via unilateral change to any other policy. However, this definition is deficient as a criterion for rational outcome because it places no constraints on the strategies players would use following histories that cannot be obtained under the given policies. These strategies do affect the NE criterion, as they enter into the opponent’s evaluation 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 \textit{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:

\[
 f^h[h'] = f[(h, h')] \hspace{1cm} (5)
 \]

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 a fixed number \( T \) of trials, with the value of \( T \) known to both players. 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). 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 possibly infinite sequence of rewards. One such criterion is the **limiting average**:

\[
V_{\text{lim}} ((r^t)_{t \geq 0}) = \lim_{r \to \infty} \frac{1}{t} \sum_{t=0}^{t-1} r^t.
\]

(6)

Another is the **total discounted reward**:

\[
V_{\text{disc}} ((r^t)_{t \geq 0}) = \sum_{t \geq 0} r^t \gamma^t.
\]

(7)

Here \(\gamma\) is the **discount factor**, with \(0 \leq \gamma < 1\). Note that the limiting average is not guaranteed to exist, while the total discounted reward is (assuming a bounded range of payoffs for the constituent game). The total discounted reward is also a convenient criterion because it corresponds to the expected total (non-discounted) reward under a finite stochastic-length game with constant continuation probability \(\gamma\), i.e. a game of length \(T\) with \(P[T \geq t] = \gamma^t\). Indeed, if \(r\) and \(\hat{r}\) denote the rewards in the infinite game and the stochastic-length game, respectively, then

\[
\sum_{t \geq 0} \mathbb{E}[\hat{r}^t] = \sum_{t \geq 0} \mathbb{E}[\hat{r}^t | \text{trial } t \text{ occurs}] \cdot P[\text{trial } t \text{ occurs}]
= \sum_{t \geq 0} \mathbb{E}[r^t \cdot \gamma^t] = \sum_{t \geq 0} \mathbb{E}[r^t] \cdot \gamma^t = E \left[ \sum_{t \geq 0} r^t \gamma^t \right] = V_{\text{disc}} (r).
\]

(8)
This correspondence is useful as it allows the more realistic stochastic-length game to be modeled using the mathematically more convenient infinite-length discounted game.

For the remainder of this paper we assume the total discounted reward for evaluating outcomes, and denote this criterion by $V$. (Nevertheless, the results presented in the following section, in particular the universality theorems, hold for the limiting average case as well.) 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

$$V_k(f_1, f_2) = E \left[ \sum_{t \geq 0} r_t^k \gamma^t \mid f_1, f_2 \right]. \quad (9)$$

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

$$V_k(f_1, f_2, h) = E \left[ \sum_{t \geq \tau} r_t^k \gamma^{t-\tau} \mid h, f_1, f_2 \right] = E \left[ \sum_{t \geq 0} r_t^k \gamma^t \mid f_1^h, f_2^h \right] = V_k(f_1^h, f_2^h). \quad (10)$$

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

$$\forall h \forall f_1 \quad V_1(f_1, f_2, h) \leq V_1(f_1^*, f_2, h). \quad (11)$$
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 (finite) set $S$ of states, an initial state $s_0$, the set of actions $A_k(s)$ available to player $k$ in state $s$, immediate rewards $r_k(a_1, a_2)$ 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 past sequence of states, as well as the present state. For technical reasons, the set $H$ of histories is restricted to those which are realizable given the transition probabilities of the game, i.e. sequences $(s^0, a_1^0, a_2^0, \ldots, s^{t-1}, a_1^{t-1}, a_2^{t-1}, s^t)$ that satisfy $p(s^{t+1} | s^\tau, a_1^\tau, a_2^\tau) > 0$ for all $\tau < t$.

An important class of policies in a stochastic game is the class $\Omega$ 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., in any history ending in $s$).

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

\[
V_k(f_1, f_2, s) = E\left[ \sum_{t \geq \tau} r_t' \gamma^{t-\tau} \mid s^\tau = s, f_1, f_2 \right]
\]

\[= V_k(f_1, f_2, h) \quad \forall h \ni c(h) = s. \tag{12} \]

The third expression here gives \( U \) explicitly as a sum of immediate and delayed rewards, and shows its dependence on \( V \).\(^1\) Conversely, the value \( V \) associated with any state can readily be seen to be equal to the expected outcome value, conditioned upon the strategies used in that state:

\[
V_k(f_1, f_2, s) = \sum_{a_1, a_2} [U_k(f_1, f_2, s, a_1, a_2) \cdot f_1[s](a_1) \cdot f_2[s](a_2)]. \tag{14}
\]

Often the policies are suppressed and we write \( V(s) \) and \( U(s, a_1, a_2) \).

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 iff neither player can gain an advantage via unilateral 1-step deviation following any history (Filar & Vrieze, 1997). Further, when both players’

\(^1\) 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).
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 which maximizes the expected value of $U$, as conditioned upon the opponent’s strategy in that state. This fact is formalized in the following result.

**Lemma 2.1 (Filar & Vrieze, 1997):** 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

$$[(U_1(f_1, f_2, s, a_1, a_2), U_2(f_1, f_2, s, a_1, a_2))]_{a_1 \in A_1(s), a_2 \in A_2(s)}.$$ \hspace{1cm} (15)

In other words, with stationary policies the SPE condition is equivalent to the requirement that in every state, players act rationally (in the sense of NE) with respect to the adjusted one-shot game in which immediate payoffs are augmented with expected delayed rewards. As is demonstrated later, this result allows explicit calculation of stationary SPEs in a stochastic game.

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. 13). 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 Proposition 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 the opponent’s past actions, which in turn forces players to consider the effects of their actions on the opponent’s 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). A best reply for the player is therefore given by any optimal policy in the MDP.² Now it is well known that the set of optimal policies in an MDP includes a pure stationary policy (assuming there are finitely many actions per state; see, e.g., Puterman, 1994, p. 154). The conclusion is summarized in the following lemma.

[Lemma 2.2 (Filar & Vrieze, 1997): 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.]

² The set of policies in the stochastic game is strictly larger than that for the MDP, because the former can take as input the opponent’s 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).
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 set of policies 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.

**Definition 3.2:** A restricted SPE within a policy class $R$ is a profile of policies in $R$ which 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:
\[ \forall h, \forall f_1 \in R: V_1(f_1, f_2^*, h) \leq V_1(f_1^*, f_2^*, h) \]
\[ \forall h, \forall f_2 \in R: V_2(f_1^*, f_2, h) \leq V_2(f_1^*, f_2^*, h). \]  

**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 \( \Omega \) of stationary policies and \( \overline{\Omega} \) 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 the to 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 \( \Omega \) and \( \overline{\Omega} \). 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 \( H \to \Xi \), where \( \Xi \) is an arbitrary set acting as the image space of the map \( I \). (It will be seen shortly that \( \Xi \) 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 \( \overline{F}_I \) be the set of pure policies within \( F_I \). As was the case for stationary policies, \( f \) is a policy for player \( k \) in \( F_I \) if and only if there exists a map \( \hat{f} : \Xi \to \bigcup_i B_k(s) \) such that \( f = \hat{f} \circ I \). Often \( f \) is identified with \( \hat{f} \), with \( f[\xi] \) representing \( f[h] \) for any \( h \) satisfying \( c(h) = \xi \). 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 \( \overline{F}_c = \overline{\Omega} \) (recall that \( c \) returns the current state of any history).

We show that \( F_I \) and \( \overline{F}_I \) are 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 $\Phi: \Xi \rightarrow S$ 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 $\Psi: \Xi \times \bigcup_s A_1(s) \times \bigcup_s A_2(s) \times S \rightarrow \Xi$ satisfying

$$\Psi(I(h), a_1, a_2, s) = I([h, a_1, a_2, s])$$

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

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 $I$ that are compatible with that state.

**Lemma 3.2:** Let $\Gamma$ 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 $(s^t)_{t \geq 0}$
is the sequence of states arising in $\Gamma$ (conditioned on the policies $f_1$ and $f_2$). Then there exists a new stochastic game $\Gamma_I$ with state-space $\Xi$, whose policies correspond one-to-one with those of $\Gamma$ in a manner that preserves expected payoffs to both players, with the sequence of resulting states and rewards given by the process $(\xi^i_t, r^i_1, r^i_2)_{t \geq 0}$.

**Proof:** The proof proceeds constructively, by explicitly defining the actions, rewards, and transition probabilities for $\Gamma_I$ and then verifying that the resulting game has the stated correspondences with $\Gamma$. Because $I$ is assumed to be sufficient, we can define $\Gamma_I$ by splitting each state $s$ of $\Gamma$ into states $\xi \in \Phi^{-1}(s)$. Actions and rewards can thereby 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 in $\Gamma_I$ by:

$$p(\xi' | \xi, a_1, a_2) = \begin{cases} p(\Phi(\xi') | \Phi(\xi), a_1, a_2) & \Psi(\xi, a_1, a_2, \Phi(\xi')) = \xi' \\ 0 & \text{otherwise} \end{cases}.$$  \hspace{1cm} (18)

Thus transition probabilities in $\Gamma_I$ are the same as in $\Gamma$, with the constraint that the succeeding $\Xi$-state $\xi'$ must correspond to the value of $I$ dictated by the prior $\Xi$-state, the actions, and the succeeding $S$-state $\Phi(\xi')$.

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

$$\Theta: (s^0, a_1^0, a_2^0, \ldots, s_i^{i-1}, a_1^{i-1}, a_2^{i-1}, s^i) \mapsto (\xi^0, a_1^0, a_2^0, \ldots, \xi_i^{i-1}, a_1^{i-1}, a_2^{i-1}, \xi^i).$$  \hspace{1cm} (19)

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

$$(\xi^0, a_1^0, a_2^0, \ldots, \xi_i^{i-1}, a_1^{i-1}, a_2^{i-1}, \xi^i) \mapsto (\Phi(\xi^0), a_1^0, a_2^0, \ldots, \Phi(\xi_i^{i-1}), a_1^{i-1}, a_2^{i-1}, \Phi(\xi^i)).$$  \hspace{1cm} (20)

Surjectivity of $\Theta$ is a result of determinism of $I$, and the fact that the transition probabilities defined for $\Gamma_I$ make the set of realizable histories in the expanded game
correspond precisely to the image of $\Theta$. (Recall that $H_{\Gamma_i}$ is defined as the set of histories that are realizable given the game’s transition probabilities.) Indeed, the condition

$$\forall \tau < t \quad \Psi(\xi^{\tau}, a_1^\tau, a_2^\tau, \Phi(\xi^{\tau+1})) = \xi^{\tau+1}$$

(21)

is necessary and 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 $\Theta$.

Finally, note that $\Theta$ also implies a bijection between policies of the two games, via $f \mapsto f \circ \Theta^{-1}$ (where $f$ is any policy in $\Gamma$). It is now straightforward to verify (by induction on $t$) that the probability distributions on the sequences of $I$-values $(I(s^t))_{t\geq0}$ and rewards $(r^t_k)_{t\geq0}$ implied by arbitrary policies $f_1$ and $f_2$ in $\Gamma$ are identical to those on $\Xi$-states $(\xi^t)_{t\geq0}$ and rewards $(r^t_k)_{t\geq0}$ implied by the corresponding policies $f_1 \circ \Theta^{-1}$ and $f_2 \circ \Theta^{-1}$ in $\Gamma_I$.

The expanded game $\Gamma_I$ 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 $\Gamma_I$ leads to novel conclusions regarding best replies and SPEs in $\Gamma$.

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

**Proof:** Observe that $F_I$ and $\overline{F}_I$ correspond precisely to the classes of stationary and pure stationary policies in the expanded game $\Gamma_I$. Indeed, a policy $f$ for $\Gamma$ is an element of $F_I$ if and only if there exists a map $\hat{f} : \Xi \to \bigcup_s B_k(s)$ with $f = \hat{f} \circ I$, whereas the corresponding policy $f \circ \Theta^{-1}$ for $\Gamma_I$ is stationary if and only if there exists a map $\tilde{f} : \Xi \to \bigcup_{\xi} B_k(\xi)$ with $f \circ \Theta^{-1} = \tilde{f} \circ c$, or equivalently, $f = \tilde{f} \circ c \circ \Theta$. Using the equivalence $\bigcup_s B_k(s) = \bigcup_\xi B_k(\xi)$ 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 \( f_2 \in F_l \) (respectively \( \bar{F}_l \)) be a policy for player 2 in \( \bar{A} \). The corresponding policy \( f_2 \circ \Theta^{-1} \) in \( \Gamma_l \) is (pure) stationary, and thus there exists a (pure) stationary best reply for player 1 by Lemma 2.2. The policy in \( \bar{A} \) that corresponds to this best reply lies in \( F_l (\bar{F}_l) \). Therefore player 1 has a best reply in \( F_l (\bar{F}_l) \), so \( F_l (\bar{F}_l) \) is best-reply universal.

**Theorem 3.4 (SPE Universality):** If \( I \) is sufficient and deterministic then \( F_l \) and \( \bar{F}_l \) are SPE universal.

Proof: This correspondence between \( F_l \) (respectively \( \bar{F}_l \)) and \( \Omega_{\Gamma_l} (\bar{\Omega}_{\Gamma_l}) \) implies that for any strategy pair \((f_1, f_2)\) that forms a restricted SPE within \( F_l (\bar{F}_l) \), the corresponding pair \((f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1})\) is an SPE within \( \Omega_{\Gamma_l} (\bar{\Omega}_{\Gamma_l}) \). Now, \( \Omega_{\Gamma_l} (\bar{\Omega}_{\Gamma_l}) \) is SPE universal in \( \Gamma_l \) by Lemma 3.1, and therefore \((f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1})\) is an SPE within the full policy space \( F_{\Gamma_l} \). Using the bijection \( \Theta \) once more we see that \((f_1, f_2)\) is an unrestricted SPE for \( \Gamma \). Therefore \( F_l \) and \( \bar{F}_l \) 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 \( \Gamma_l \).

**Proposition 3.5:** Assume \( I \) is sufficient and deterministic. Then a pair of policies \( f_1, f_2 \in F_l \) forms an SPE if and only if for every value \( \xi \) of \( I \) the strategies \( f_1[\xi] \) and \( f_2[\xi] \) constitute an NE for the one-shot \( U^I \)-game with payoffs given by

\[
\left[ \left( U^I_1 (f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1}, \xi, a_1, a_2), U^I_2 (f_1 \circ \Theta^{-1}, f_2 \circ \Theta^{-1}, \xi, a_1, a_2) \right) \right]_{a_1 \in A_1(\Phi(\xi)), a_2 \in A_2(\Phi(\xi))}. \tag{22}
\]

Proof: Apply Lemma 2.1 to the policies \( f_1 \circ \Theta^{-1} \) and \( f_2 \circ \Theta^{-1} \) in \( \Gamma_l \). 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. Furthermore, while rational analysis of the one-shot game dictates that a selfish rational agent will always defect, as this strategy is dominant, behavioral results contradict this prediction (Rapoport & Chammah, 1965). It has long been known that a rational basis for cooperation can arise when 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 IPD 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.\(^3\) 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 \( \Gamma_i \), which has four states \( \xi \) corresponding to the values of \( I \), with transition probabilities dependent only on the actions: \( \forall \xi \, p(CC \mid \xi, C, C) = 1, \) etc. Stationary policies in this game are functions

\(^3\) This definition does not determine the value of \( I(\emptyset) \), 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.
that specify a probability of cooperating in each state, 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 \( \xi \).

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 \( \Gamma_0 \)), 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 sets: 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 the set of pure stationary policies, and check for pairs that are mutual best replies. In the second case the analysis relies on the result of Proposition 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 \( \Gamma_I \) 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 \( \hat{A}_I \).) 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 constant “\( U \)-shape,” and to systematically determine the set of symmetric SPEs for each region.

**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. The SPE universality theorem as applied to pure policies ensures that any such match will be an unrestricted SPE. These SPEs form the boundary points of the SPE manifold in 8-dimensional joint policy space.
Because of the subgame-perfect criterion, a best reply for player 1 to any policy of player 2 must simultaneously maximize $V_1(\xi)$ for all four values of $\xi$ (see Definition 2.1). In principle all 16 pure policies must be tested for satisfaction of this criterion, although usually groups can be eliminated at once, such as any policy $f_1$ for player 1 with $f_1[CD] = 1$ when $f_2[CD] = 0$ (this combination leads to $V_1(CD) = x/(1-\gamma) < 0$, while choosing the all-defect policy $f_1[\xi] \equiv 0$ would guarantee all state values for player 1 to be nonnegative). 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 $\Gamma_1$. Best-reply universality implies that each of these is a universally best reply, without restriction to 1-back memory.

//Insert Table 1 about here//

For sufficiently small $\gamma$ the best reply to any of these policies 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 $\gamma$. However, there are some policies for which the best reply involves cooperation, implying that if the opponent were using such a strategy then it would be rationally justified to cooperate in certain states. TFT for player 2 $- \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ – induces the most cooperation from player 1, in that it is the only policy for which all-cooperate is a best reply, and that it induces a semi-cooperative best reply for the smallest value of $\gamma$ (aside from “exploitable” policies having $f_2[DC] = 1$). However, TFT as a policy for player 1 $- \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ – is only a best reply when player 2 uses
either \[
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\] or TFT under a precise \(\gamma\) value; in these situations, half or all of the pure policies are best replies (i.e., at least three components of player 1’s policy are irrelevant to his or her expected total reward).

The tabulation of best replies given in Table 1 now allows determination of the set of pure stationary SPEs for \(\Gamma_i\) (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 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 (#1) is the Grim strategy (Friedman, 1971), which can cooperate indefinitely but responds to any deviation with unending defection.

Equilibrium #2 involves a similar strategy 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 defection or to a sequence in which the two players alternate with one 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 \(\gamma\); we will see, however, that there is a whole family of SPEs in mixed policy space that jointly exist for a range of \(\gamma\)-values, of which the present one is a boundary point. SPE #4 is a minor variation of #3, and is asymmetric (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 Tit-For-Tat strategy makes the other player’s entire policy irrelevant to their 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.

//Insert Table 2 about here//

Symmetric SPEs

To deal with the general case of mixed strategies, we make use of the relation
between SPEs and U-game NEs implied by Lemma 2.1 and Proposition 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.
However, because transition probabilities in $\Gamma_i$ are independent of the present state,
expected future rewards (assuming stationary policies) and therefore the components of
the $U$-game are also independent of state. Another way to see this is that the state on trial
$n$ corresponds to the actions in trial $n - 1$, while the state on trial $n + 1$ corresponds to the
actions in trial $n$; hence state $n + 1$ is independent of state $n$ (when conditioned upon the
actions in trial $n$). Formally, for any states $\xi$ and $\zeta$, stationary policies $f_1$ and $f_2$, and
actions $a_1$ and $a_2$, we have

$$U_k (f_1, f_2, \xi, a_1, a_2) = r_k (a_1, a_2) + \gamma \cdot \sum_{\xi'} [V_k (f_1, f_2, \xi') \cdot p(\xi' | \xi, a_1, a_2)]$$

(22)

$$= r_k (a_1, a_2) + \gamma \cdot \sum_{\xi'} [V_k (f_1, f_2, \xi') \cdot p(\xi' | \xi, a_1, a_2)] = U_k (f_1, f_2, \zeta, a_1, a_2).$$

Because of this state invariance we drop the state $\xi$, 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. This in turn leads the $U$-game to be symmetric as well, that is

\[
U_1(C, C) = U_2(C, C) \quad U_1(C, D) = U_2(D, C) \\
U_1(D, C) = U_2(C, D) \quad U_1(D, D) = U_2(D, D).
\]

(23)

The next step is to characterize the possible NEs for this $U$-game. In general, the NEs of a $2 \times 2$ 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 Figure 1. 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.

//Insert Figure 1 about here//

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

\[
U_1(C, D) \preceq U_1(D, D) \quad U_1(C, C) \preceq U_1(D, C)
\]

(24)

Each of these can take truth values of ‘$>$’, ‘$<$’, or ‘$=$’. Under the symmetry assumption, these comparisons determine corresponding comparisons for $U_2$, by (23). The best-reply graph in Figure 1 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. Figure 2 shows the range of expected joint payoffs to the two players under all possible strategies, for a reward matrix consistent with the example of Figure 1. 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 other relations than just those in (24), these relations 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.

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 Figure 3. 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 (13), we see that each component
of \( U \) is offset from the corresponding \( r \) by a scaled-down convex combination of the \( U \)s
themselves:

\[
U(a_1, a_2) = r(a_1, a_2) + \gamma \cdot (f_1[a_1a_2]f_2[a_1a_2] \cdot U(C, C) + f_1[a_1a_2](1 - f_2[a_1a_2]) \cdot U(C, D) + (1 - f_1[a_1a_2])f_2[a_1a_2] \cdot U(D, C) + (1 - f_1[a_1a_2])(1 - f_2[a_1a_2]) \cdot U(D, D)).
\]  \(\text{(23)}\)

(Here \( f_k[a_1a_2] \) represents player \( k \)’s probability of cooperating in state “\( a_1a_2 \)” i.e.,
following an outcome of \((a_1, a_2)\).) Figure 4 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
(24) and thus changing the best-reply graph to the form shown in Figure 1, which
includes 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 (24), each with a qualitatively different set of NEs (see 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. What remains is to determine which of these sets of assignments are truly SPEs, in that the $U$-games induced by the policies have the structure originally assumed (so that the strategy pairs in 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. 12), 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 $\gamma$; the specific relationships are given in Appendix C.

All of the first nine SPEs listed in Table 4 (including the following two duplicates) are repetitions or generalizations of deterministic equilibria #1 and #2 shown in Table 2, in that they can be put into the form of one of the 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 $\begin{pmatrix} p & q \\ r & s \end{pmatrix}$ vs. $\begin{pmatrix} p & r \\ q & s \end{pmatrix}$ is a generalization of deterministic SPE #5, in that the four action probabilities have the property of collectively making the opponent’s entire 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.

**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, the idea of one type of policy arising from another is somewhat problematic. However, this sort of scenario does make sense from the perspective of dual timescales, 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 $\gamma$ 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.

In order to determine the connectedness of all-defect to the other SPEs derived here, we need only consider those SPEs for which no component of either player’s policy is equal to 1, that is all action probabilities are 0 or mixed. The candidate equilibria are therefore $\begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix}$ vs. $\begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}$ vs. $\begin{pmatrix} 0 & 0 \\ 0 & p \end{pmatrix}$ vs. $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, and $\begin{pmatrix} p & q \\ r & s \end{pmatrix}$ vs.
What is required is to determine the conditions under which the action probabilities in any of these equilibria all tend to 0. Using the relations in Appendix C, it can be shown that in no case do any of these SPEs become equal to the all-defect pair under the strict requirements on the payoff parameters $x < 0$ and $y > 1$ (even in the limit $\gamma \to 1$). Therefore all-defect cannot be reached by continuous variation from any semi-cooperative SPE. In other words, all-defect is an isolated path-connected component of the SPE manifold (at least when we restrict to symmetric SPEs defined by 1-back policies, as has been assumed here).

In order to determine whether all-defect is truly an isolated solution, we must use the looser requirement of connectedness, rather than path-connectedness. For this, it is necessary to consider the continuation of the solutions onto the boundary of the parameter space, by allowing $x \to 0$, $y \to 1$, or both. When both of these limits are satisfied, and $\gamma = 0$, the IPD becomes a one-shot game in which neither player’s action affects his or her own payoff, and thus every policy pair is an SPE. This implies that in the closed parameter space (the proper parameter space plus boundary), the solution branches corresponding to all-defect and any other SPE are connected. However, the relevant question is whether it is possible to move continuously away from all-defect while the game changes from $(x, y) = (0, 1)$ to a proper Prisoner’s Dilemma, or equivalently whether there exist SPEs arbitrarily close to all-defect within the proper parameter space (this latter condition is the meaning of connectedness in this context). In the case of $\begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ 0 & p \end{bmatrix}$, it can be shown that $p \geq \frac{1}{2}$ whenever $(x, y, \gamma) \neq (0, 1, 0)$, and thus this SPE is strictly bounded away from all-defect within the proper parameter space. This situation is illustrated schematically in Figure 5A. In contrast, for $\begin{bmatrix} 0 & 0 \\ p & 0 \end{bmatrix}$ vs. $\begin{bmatrix} 0 & p \\ 0 & 0 \end{bmatrix}$, we have $\lim_{x \to 0} (p) = 0$ for any $y \geq 1$ and $\gamma \in (0, 1)$. Thus allowing $x$ to approach 0 leads to semi-cooperative SPEs arbitrarily close to all-defect (see Figure 5B).
A similar situation arises for $\begin{bmatrix} p & 0 \\ 0 & 0 \end{bmatrix}$ vs. $\begin{bmatrix} p & 0 \\ r & s \end{bmatrix}$ 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 $(x, y)$ tend jointly to $(0, 1)$.\footnote{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.} Therefore these latter three families of SPEs all provide a potential explanation for the emergence of cooperation, under the dual timescale interpretation used here.

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

Perhaps the most important general conclusion to take away from the universality theorems concerns the information people do and do not use. Intuitively, the best-reply universality theorem states that if one’s opponent ignores certain publicly available information in making his or her decisions, then one cannot benefit from using this information oneself (provided it provides no additional knowledge of the current state of the game; this is the import of the sufficiency axiom). A converse to this is that if the
opponent is basing actions on certain information, even if it is entirely unrelated to the workings of the game, then the focal player may be able to benefit from using that information as well. A simple example is a 2×2 pure coordination game, in which the players both receive an equal reward if they manage to select matching actions. If there happens to be some commonly observable (though ostensibly unrelated) process taking place, such as a bystander repeatedly flipping a coin, then there exist SPEs in which players choose actions according to the most recent outcome of this process. By both using the information of the extraneous stimulus according to some arbitrary convention, the players can coordinate their actions and thus improve their payoffs.

From a theoretical standpoint, the SPE universality theorem provides a powerful tool for evaluating the SPE criterion for a variety of types of policies. By giving a simple criterion for subclasses of policies to be SPE universal, the theorem reduces the equilibrium requirement to one of optimality with respect to a much smaller set of alternatives. Further, the method shown here allows for the discovery of SPEs in an arbitrary repeated or stochastic game, often composed of simple and interpretable policies, through various (perhaps independently motivated) choices of the function I. Discovery in this way of new theoretical rationally supported behaviors may then have implications for understanding realistic situations.

All of these points are well illustrated by our investigation of mutually optimal 1-back policies in the IPD. In this game it was seen that ignorance of the game history further than one trial back allows players to use cognitively simple policies that are nevertheless optimal with respect to each other, and which are able to achieve maximally cooperative outcomes. Classification and verification of these SPEs was greatly facilitated by the restriction to stationary policies in the 4-state expanded game, as justified by the SPE universality theorem, and by the U-game partitioning technique. The class of policies thus discovered includes some that are well known, such as Grim and
TFT, but also includes a number of others that can manifest in a variety of interesting semi-cooperative joint behaviors.

Some particularly interesting results arose regarding TFT, which is rarely a best reply to any policy although the best replies to it can be highly cooperative. The reason for this seeming contradiction – TFT is effective yet is rarely a best reply – is that the power of TFT is restricted to situations in which the opponent is able to assess the contingency between his actions and TFT’s subsequent replies. This is the case when we consider the set of best replies to TFT, as what is really being modeled is an opponent who is fully aware of the policy that he or she is playing against and is selecting a policy that will play well against it. In this situation the power of TFT is revealed in that the policy chosen will often be highly cooperative. However, when determining the best reply to a policy of the opponent, the situation is quite different. In this case the opponent’s policy is fixed, and TFT is only especially useful if that policy is one which has some ability to learn the future effects of its own actions. In cases where the opponent’s policy is as simple as one with 1-back memory, TFT will not be especially useful, as compared to other policies that go along with the “stubbornness” of the opponent’s policy rather than trying to “teach” it.

Of course, any process by which the opponent identifies the player’s policy during play and learns to respond to it optimally could 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 very recent information). A consequence of the best-reply universality theorem is that if the opponent’s policy is temporally bounded, then this 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 ignores the need for online learning by essentially assuming that players know each other’s policy beforehand. 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 in repeated interactions as a consequence of online learning. 

SPE Universality and Machine Games

The concepts of restricted SPEs and informationally bounded policies discussed here have close connections to 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, Ch. 9). In a machine game, each player chooses actions based on the state of some automaton whose updating process depends on the outcomes of a repeated matrix game. 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.

---

5 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 opponents’ policy. In the absence of this assumption more complex algorithms could well fare better.
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. 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 is known to exist a machine-based deterministic SPE that produces the same sequence of outcomes (Osborne & Rubinstein, 1994).

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 result can be interpreted as stating that the class $M_n$ 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 the 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 result can therefore be viewed as an extension of Abreu and Rubinstein’s (1988) theorem, which generalizes it in a number of ways. First, the present

---

6 Indeed, if an automaton $A_I$ is defined with $Ξ$ as its set of states and $Ψ$ as its transition function, then when allowed to run in synchrony with the play of a repeated game it will at every step be in the state corresponding to the current value of $I$. Thus the classes $F_I$ and $F_I$ of mixed and pure policies based on $I$ are precisely the sets $M(A_I)$ and $M(A_I)$ of mixed and pure policies that can be implemented based on the automaton. Technically, all SPEs considered here must be implemented using the same automaton (i.e., same information $I$); however, given any two policies each based on a different automaton, one can construct the product of the automata and base both policies on that machine.

7 In general, it can be easily shown that any class that is a union of best-reply universal classes is itself universal. In this case $M_n = \bigcup_{I \in Ω_n} F_I$, where $Ω_n$ is the set of deterministic $I$ that take on no more than $n$ values.
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 don’t 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 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 theorem does not apply. 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 earlier 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 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 property holds, namely the sufficiency axiom. This condition is trivial in a repeated game, but in a stochastic game is it critical to ensuring the SPE criterion is met, as the following example illustrates. Consider a 2-player stochastic game with two states, for which the transition probabilities are always .5 regardless of players’ actions. If $I$ is a constant function, then $I$ is deterministic but not sufficient, and furthermore any policy based on $I$ will treat the game as a one-shot game with payoffs given as the average of the payoffs in the two states. Thus any restricted SPE with respect to $F_I$ is a repetition of an NE of this averaged one-shot game. But if player 1 has opposing dominant strategies in the two states, then knowledge of the current state would allow for increased payoffs (note that since player 2’s policy is stationary, the value to player 1 of a 1-step deviation based on knowledge of the present state is precisely equal to the increase in immediate reward). Thus the restricted $F_I$-SPE is not an unrestricted SPE, and the best reply within $F_I$ is not universally optimal.

The importance of the sufficiency axiom can also be seen in light of the fact that bounded complexity of policies can lead to cooperation in the finitely-repeated Prisoner’s Dilemma (Neyman, 1985). More precisely, this result states that there exist cooperative policies that comprise SPEs when restricted to $M_n$ for $n$ sufficiently small relative to the length of the game. Because the backward induction argument demonstrates that there are no SPEs other than all-defect in a finite Prisoner’s Dilemma (Luce & Raiffa, 1957), $M_n$ cannot be 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 why the universality properties do not extend to the finite game.
References


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 $\gamma$.

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:

$$V_1(CC) = V_1(DD) = \frac{y}{1-\gamma^2}.$$  \hspace{1cm} (A1)

Similar direct calculations yield

$$V_1(CD) = V_1(DC) = \frac{\gamma y}{1-\gamma^2}.$$  \hspace{1cm} (A2)

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 she were to cooperate, the she would receive a reward of $x$ (as the opponent would be defecting), and the game would remain in $CD$ indefinitely. Thus we would have

$$V_1(CD) = \frac{x}{1-\gamma} < \frac{\gamma y}{1-\gamma^2}.$$  \hspace{1cm} (A3)
which would be inferior to all-defect. Therefore the optimal policy for the present case has \( f_1[CD] = 0 \). Now consider \( f_1[DC] \): If \( f_1[DC] = 1 \), then \( DC \) would always lead to \( CD \) followed by \( DD \), whereas if \( f_1[DC] = 0 \) then \( DC \) would lead directly to \( DD \). Thus:

\[
\begin{align*}
f_1[DC] = 1 & \Rightarrow V_1(DC) = x + \gamma^2 V_1(DD) \\
f_1[DC] = 0 & \Rightarrow V_1(DC) = \gamma V_1(DD).
\end{align*}
\] (A4)

By (A1), the optimal policy satisfies \( V_1(DD) > 0 \), and this along with \( x < 0 \) and \( 0 \leq \gamma < 1 \) implies that the second expression for \( V_1(DC) \) in (A4) is strictly greater than the first. Therefore \( f_1[DC] = 0 \). This also implies \( V_1(DD) = \gamma V_1(DD) \), so maximization of \( V_1(DD) \) will ensure that of \( V_1(BC) \). It can be similarly shown that maximizing \( V_1(DD) \) also leads to maximization of \( V_1(BC) \).

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:

\[
\begin{align*}
f_1[CC] = 0, f_1[DD] = 0 & \quad \Rightarrow \quad V_1(CC) = \frac{y}{1-\gamma^2}, \quad V_1(DD) = \frac{y}{1-\gamma^3} \\
f_1[CC] = 0, f_1[DD] = 1 & \quad \Rightarrow \quad V_1(CC) = \frac{y + \gamma^2}{1-\gamma}, \quad V_1(DD) = \frac{1 + \gamma y}{1-\gamma^3} \\
f_1[CC] = 1, f_1[DD] = 0 & \quad \Rightarrow \quad V_1(CC) = \frac{1}{1-\gamma}, \quad V_1(DD) = \frac{\gamma}{1-\gamma^2} \\
f_1[CC] = 1, f_1[DD] = 1 & \quad \Rightarrow \quad V_1(CC) = \frac{1}{1-\gamma}, \quad V_1(DD) = \frac{1}{1-\gamma}.
\end{align*}
\] (A5)

When \( \gamma < y - 1 \), the potential state values in (A5) are ordered as follows:

\[
\frac{y}{1-\gamma^2} > \frac{y + \gamma^2}{1-\gamma} > \frac{1 + \gamma y}{1-\gamma^3} > \frac{1}{1-\gamma}.
\] (A6)
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 – $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, 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:

$$f = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow U_1 = \begin{bmatrix} 1 + \frac{\gamma y}{1 - \gamma^2} & \frac{x + \gamma^2 y}{1 - \gamma^2} \\ \frac{y}{1 - \gamma^2} & \frac{\gamma y}{1 - \gamma^2} \end{bmatrix}$$

$$f = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow U_1 = \begin{bmatrix} 1 + \frac{\gamma y}{1 - \gamma^3} & \frac{x + \gamma^2 + \gamma^3 y}{1 - \gamma^3} \\ \frac{y + \gamma^2}{1 - \gamma^3} & \frac{\gamma + \gamma^2 y}{1 - \gamma^3} \end{bmatrix}$$

(A7)

$$f = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow U_1 = \begin{bmatrix} \frac{1}{1 - \gamma} & \frac{x + \gamma^2 y}{1 - \gamma^2} \\ \frac{y}{1 - \gamma^2} & \frac{\gamma y}{1 - \gamma^2} \end{bmatrix}$$

$$f = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow U_1 = \begin{bmatrix} \frac{1}{1 - \gamma} & \frac{x + \gamma^2}{1 - \gamma} \\ \frac{y + \gamma^2}{1 - \gamma} & \frac{\gamma}{1 - \gamma} \end{bmatrix}.$$
\( \gamma \) (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

$$p = \frac{U_1(D, D) - U_1(C, D)}{U_1(C, C) - U_1(D, C) - U_1(C, D) + U_1(D, D)}$$  \hspace{1cm} (B1)$$

Therefore we need to search through all policy pairs $(f, f^T)$ satisfying $(f[\xi], f^T[\xi]) = (0, 0), (1, 1), \text{or } (p, p)$ for every state $\xi$. Note that the symmetry present in all three NEs, along with the symmetry assumed between the players’ policies, implies $f[CD] = f^T[CD] = f[DC]$. 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 (13):

$$U_1(a_1, a_2) = r_1(a_1, a_2) + \gamma V_1(a_1a_2).$$ \hspace{1cm} (B2)$$

(Here “$a_1a_2$” indicates the state following the action pair $(a_1, a_2)$, as with, e.g., $CC$.) In particular, given that $r_1(C, C) < r_1(D, C)$, the assumption $U_1(C, C) > U_1(D, C)$ implies

$$V_1(CC) > V_1(DC).$$  \hspace{1cm} (B3)$$

Furthermore, the symmetry assumed for $U$ and for the policies implies both $V_1(\xi) = V_2(\xi)$ for every $\xi$ and $V_k(CD) = V_k(DC)$ (see Eq. 14). 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 $b = 0, 1, \text{ or } p$, define the equilibrium value $W^{(b, b)}$ by

$$W^{(b, b)} = \mathbb{E}\left[ \sum_{j \in \mathcal{J}} \gamma^{t-\tau^j} r^j_t \mid P[a_1 = C] = P[a_2 = C] = b \right]. \quad (B4)$$

Note that $W^{(b, b)}$ is equal to the state value $V(\xi)$ for any $\xi$ for which $(f(\xi), f^T(\xi)) = (b, b)$ (see Eq. 12). 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 (14) 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 $(b, b)$. Direct calculations yield:

$$W^{(1,1)} = U_1(C, C)$$
$$W^{(0,0)} = U_1(D, D)$$
$$W^{(p,p)} = p^2 \cdot U_1(C, C) + p(1 - p) \cdot U_1(C, D) + p(1 - p) \cdot U_1(D, C) + (1 - p)^2 \cdot U_1(D, D). \quad (B5)$$

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:

$$W^{(p,p)} = p \cdot U_1(C, C) + (1 - p) \cdot U_1(C, D) = p \cdot U_1(D, C) + (1 - p) \cdot U_1(D, D). \quad (B6)$$
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 \( W_{(p,p)} \geq U_1(C, C) \). By the first equality in (B6) this implies \( U_1(C, C) \leq U_1(C, D) \). 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 \( U_1(D, D) > U_1(C, C) > U_1(D, C) \). The second equality in Equation (B6) (LHS = RHS) now yields \( W_{(p,p)} < U_1(D, D) \). Therefore, referring back to the hypothetical assumption \( W_{(p,p)} \geq U_1(C, C) \), we can conclude that either \( W_{(p,p)} < U_1(C, C) \) or \( W_{(p,p)} < U_1(D, D) \). By (B5) this is equivalent to:

\[
W_{(p,p)} < W_{(0,0)} \text{ or } W_{(p,p)} < W_{(1,1)}.
\]  (B7)

The possible orderings of the \( W \)s consistent with (B7) are broken into three cases:

Case 1: \( W_{(1,1)} > W_{(p,p)} \geq W_{(0,0)} \)

In this case, (B3) 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:

\[
\begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & p \\ p & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & p \\ p & p \end{pmatrix}, \quad \begin{pmatrix} 1 & p \\ p & 0 \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & p \\ p & 0 \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 0 & 1 \\ 0 & p \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 0 & 1 \\ 0 & p \end{pmatrix}, \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} \text{ vs. } \begin{pmatrix} 0 & 1 \\ 0 & p \end{pmatrix}.
\]
Case 2: \( W^{(1, 1)} > W^{(0, 0)} > W^{(p, p)} \)

In this case the only way to satisfy (B3) 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 (B6) 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 (B2). Therefore \( V(DD) \neq W^{(p, p)} \), implying \( f[\xi] \neq p \). This leaves two new possibilities:
\[
\begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix} \text{ vs. } \begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 \\ p & 1 \end{bmatrix} \text{ vs. } \begin{bmatrix} 0 & p \\ p & 1 \end{bmatrix}.
\]

Case 3: \( W^{(0, 0)} > W^{(p, p)}, W^{(0, 0)} \geq W^{(1, 1)} \)

The assumption \( W^{(0, 0)} \geq W^{(1, 1)} \) is equivalent to \( U(D, D) \geq U(C, C) \) by (B5). Using \( r_1(D, D) < r_1(C, C) \), (B2) implies \( V(DD) > V(CC) \). Taking into account (B3) 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 \( \begin{bmatrix} p & 1 \\ 1 & 0 \end{bmatrix} \text{ vs. } \begin{bmatrix} p & 1 \\ 1 & 0 \end{bmatrix} \).

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

With this policy pair mutual cooperation perpetuates itself indefinitely, so we see readily that
\[ U_1(C, C) = \frac{1}{1 - \gamma}. \]  
(B8)

Using Equations (B2) and (B8) to substitute for the components of \( U \) in Equation (B6), 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:

\[ W^{(p, p)} = \frac{p}{1 - \gamma} + (1 - p) \cdot (x + \gamma \cdot W^{(p, p)}) = p \cdot (y + \gamma \cdot W^{(p, p)}) + (1 - p) \cdot \gamma \cdot W^{(p, p)} \]  
(B9)

\[ \Rightarrow W^{(p, p)} = \frac{p + (1 - p)(1 - \gamma) x}{(1 - \gamma)(1 - \gamma + \gamma p)} \quad \text{and} \quad W^{(p, p)} = \frac{p y}{(1 - \gamma)} \]  
(B10)

\[ \Rightarrow \gamma p^2 + ((1 - \gamma)(x + y) - 1)p - (1 - \gamma)x = 0. \]  
(B11)

Under the restriction to \(0 \leq p \leq 1\) and \(0 \leq \gamma < 1\), then there is one solution curve, given by

\[ p = \frac{1 - (1 - \gamma)(x + y) \pm \sqrt{(1 - \gamma)^2 (x^2 + y^2) + 2(1 - \gamma)^2 xy - 2(1 - \gamma)(x + y) + 1}}{2 \gamma} \]  
(B12)

with boundary points at \((\gamma, p) = (1, 0)\) and \((1, 1/y)\). This curve represents the set of pairs \((\gamma, p)\) 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 \(U_1(C, C) \geq U_1(D, C)\). On the solution curve given by (B12), where \((p, p)\) is an NE, this is equivalent to the requirement \(U_1(C, D) \leq U_1(D, D)\) by (B6). Using (B2), we have
\[ U_1(C, D) = x + \gamma V(CD) = x + \gamma W^{(p, p)} \]  \hspace{1cm} (B13)

and

\[ U_1(D, D) = 0 + \gamma V(DD) = \gamma W^{(p, p)}. \]  \hspace{1cm} (B14)

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 (B12). Therefore \[
\begin{bmatrix}
1 & p \\
p & p
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
1 & p \\
p & p
\end{bmatrix}
\] is an SPE for all \( \gamma \) such that the values of \( p \) given by (B12) are real and lie in \([0, 1]\). It can be shown algebraically that this condition is satisfied whenever \( \gamma \geq \gamma^* \), with the critical value \( \gamma^* \in (0, 1) \) given by

\[ \gamma^* = \frac{x(x-1) + y(y-1) + 2\sqrt{xy(x-1)(y-1)}}{(x-y)^2}. \]  \hspace{1cm} (B15)

\[
\begin{bmatrix}
0 & p \\
p & 0
\end{bmatrix}
\] 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 \( U_1(C, D) \leq U_1(D, D) \). This along with (B6) (the criterion for \( (p, p) \) to be an NE) implies \( U_1(D, C) \leq U_1(C, C) \). Since state \( CC \) is followed by eternal defection, \( U_1(C, C) = 1 \). Therefore by (B2):

\[ V(DC) = \frac{1}{\gamma} (U(D, C) - y) \leq \frac{1}{\gamma} (U(C, C) - y) = \frac{1}{\gamma} (1 - y) < 0. \]  \hspace{1cm} (B16)

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 \( \gamma \geq 1 - 1/y \).

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

Exists iff \( y < 2 \). Requires \( \gamma \geq y - 1 \).

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

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

Exists \( \forall x, y \). Requires \( \gamma \geq 1 - 1/y \).

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

\[
p = \frac{1}{2(1 - \gamma)(1 - \gamma)x + (1 - \gamma)y} \cdot \left[ 1 - (1 - \gamma^2)x - (1 - \gamma^2)y \pm \sqrt{(1 - \gamma)^4(x^2 + y^2) + 2(1 - \gamma)^2(1 + 2\gamma - \gamma^2)xy - 2(1 - \gamma)^2(x + y) + 1} \right]
\]

and \( p \leq \frac{-x}{y - x - 1} \).

Exists \( \forall x, y \).
\[
\begin{bmatrix}
1 & p \\
p & p
\end{bmatrix}
\quad \text{vs.} \quad 
\begin{bmatrix}
1 & p \\
p & p
\end{bmatrix}
\]

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

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

\[
\begin{bmatrix}
1 & p \\
p & 1
\end{bmatrix}
\quad \text{vs.} \quad 
\begin{bmatrix}
1 & p \\
p & 1
\end{bmatrix}
\]

\[
p = \left[\gamma(y - x + 1)\right]^{-1} \cdot \left[1 + 2\gamma - (1 + \gamma)x - (1 - \gamma)y \right.
\]
\[
\pm \sqrt{(1 - \gamma)^2(x^2 + y^2) + 2(1 + 2\gamma - \gamma^2)xy - 2(1 + \gamma)(x + y) + 1 + 4\gamma} \right].
\]

Exists iff \( y \leq 1 + \frac{1}{4(1 - x)} \).

\[
\begin{bmatrix}
1 & 0 \\
0 & p
\end{bmatrix}
\quad \text{vs.} \quad 
\begin{bmatrix}
1 & 0 \\
0 & p
\end{bmatrix}
\]

\[
p = \left[2\gamma(1 - (1 - \gamma)x - \gamma y)\right]^{-1} \cdot \left[(1 - \gamma)^2x + (1 - \gamma^2)y - 1 \right.
\]
\[
- \sqrt{(1 - \gamma^2)^2(x^2 + y^2) + 2(1 - \gamma)(1 - \gamma + \gamma^2 + \gamma^3)xy - 2(1 - \gamma^2)(x + y) + 1} \right].
\]

Exists \( \forall x, y \).

\[
\begin{bmatrix}
p & 0 \\
0 & p
\end{bmatrix}
\quad \text{vs.} \quad 
\begin{bmatrix}
p & 0 \\
0 & p
\end{bmatrix}
\]

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

Exists iff \( y < 2 \). Requires \( \gamma \geq y - 1 \).

54
$$\begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix} \text{ vs. } \begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix}$$

$$p = [2\gamma(-\gamma a + (1+\gamma)b)]^{-1} \left[ (1+\gamma^2)a + (1+\gamma^2)b - 1 - 2\gamma \\
+ \sqrt{(1+\gamma^2)^2(a^2 + b^2) + 2(1-5\gamma^2 - 2\gamma^2 - \gamma^4)ab - 2(1+3\gamma^2)(a+b) + (1+2\gamma)^2} \right].$$

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

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

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

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

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

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

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

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

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

Requires $\frac{\gamma s + 1)x + \gamma(1-s)y}{\gamma x} \leq q \leq \frac{(-\gamma s - 1 + \gamma)y + 1 + \gamma s}{\gamma}$, which requires $\gamma \geq \max\{1 - \frac{1}{y}, 1 - \frac{1}{1-x}\}$. Exist solutions $\forall x, y$. 

55
\[
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\]

Exists \( \forall x, y, \gamma \).
Author Note

Matt Jones, Department of Psychology; Jun Zhang, Department of Psychology.

This work was supported in part by a Pre-doctoral Fellowship from the Horace H. Rackham Graduate School at the University of Michigan.

Correspondence concerning this article should be addressed to Matt Jones, who is now at the Department of Psychology, University of Texas at Austin, 1 University Station A8000, Austin, Texas 78712. Email: mattj@psy.utexas.edu
Table 1. Best replies to all pure 1-back policies in the IPD.

<table>
<thead>
<tr>
<th>Policy (Player 2)</th>
<th>Best Replies (Player 1)</th>
<th>Discounting Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>all ( \gamma )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 1 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>all ( \gamma )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 1 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>all ( \gamma )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 1 &amp; 1 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>all ( \gamma )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 1 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 1 &amp; 0 \ 1 &amp; 1 \end{bmatrix} )</td>
<td>( -\frac{x}{y} \leq \gamma &lt; 1 )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 1 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 1 &amp; 0 \end{bmatrix} )</td>
<td>( \gamma = -\frac{x}{y} )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( 0 \leq \gamma \leq -\frac{x}{y} )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 1 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>all ( \gamma )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 1 \ 1 &amp; 1 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix} )</td>
<td>( -\frac{x}{y-x} \leq \gamma &lt; 1 )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 1 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \gamma = -\frac{x}{y-x} )</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( \begin{bmatrix} 0 &amp; 0 \ 0 &amp; 0 \end{bmatrix} )</td>
<td>( 0 \leq \gamma \leq -\frac{x}{y-x} )</td>
</tr>
</tbody>
</table>

(Table 1 continued on following page)
\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\quad 1 - \frac{1}{y} \leq \gamma < 1
\]

\[
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\quad y - 1 \leq \gamma < 1
\]

\[
\begin{bmatrix}
1 & 0 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\quad 0 \leq \gamma \leq 1 - \frac{1}{y}
\]

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\quad \gamma = y - 1
\]

\[
\begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\quad 0 \leq \gamma \leq y - 1
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
1 & 1
\end{bmatrix}
\quad \text{all } \gamma
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\quad \text{all } \gamma
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
1 & 1
\end{bmatrix}
\quad \max \left\{ \frac{y - 1}{1 - x}, \frac{x}{1 - x} \right\} \leq \gamma < 1
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\quad \gamma = -\frac{x}{1 - x}, \text{ with } x + y = 1
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\quad \frac{y - 1}{x} \geq -\frac{x}{1 - x}
\]

\[
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\quad 0 \leq \gamma \leq \min \left\{ 1 - \frac{x}{y}, -\frac{x}{y} \right\}
\]

(Table 1 continued on following page)
### BOUNDED INFORMATION

\[
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix} \quad y - 1 \leq \gamma < 1
\]

\[
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 0
\end{bmatrix} \quad \gamma = y - 1
\]

\[
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix} \quad 0 \leq \gamma \leq y - 1
\]

\[
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix} \quad - \frac{x}{y - x} \leq \gamma < 1
\]

\[
\begin{bmatrix}
1 & 1 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix} \quad 0 \leq \gamma \leq - \frac{x}{y - x}
\]

\[
\begin{bmatrix}
1 & 1 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix} \quad \text{all } \gamma
\]

**Notes:** Discounting range gives constraints on \( \gamma \) for the given reply(s) to be optimal. The ranges for \( \gamma \) depend on the payoff parameters \( x \) and \( y \); not all cases will occur for all values of these parameters.
Table 2. Pure 1-back SPEs for the IPD.

<table>
<thead>
<tr>
<th>Equilibrium Policies</th>
<th>Player 1</th>
<th>Player 2</th>
<th>Discounting Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>[1 0]</code></td>
<td><code>[1 0]</code></td>
<td></td>
<td>( 1 - \frac{1}{y} \leq \gamma &lt; 1 )</td>
</tr>
<tr>
<td><code>[0 0]</code></td>
<td><code>[0 0]</code></td>
<td></td>
<td>( y - 1 \leq \gamma &lt; 1 )</td>
</tr>
<tr>
<td><code>[1 0]</code></td>
<td><code>[1 0]</code></td>
<td></td>
<td>( \gamma = -\frac{x}{y} )</td>
</tr>
<tr>
<td><code>[0 1]</code></td>
<td><code>[0 1]</code></td>
<td></td>
<td>( \gamma = -\frac{x}{y} )</td>
</tr>
<tr>
<td><code>[1 0]</code></td>
<td><code>[0 0]</code></td>
<td></td>
<td>( \gamma = -\frac{x}{y} ), with ( x + y = 1 )</td>
</tr>
<tr>
<td><code>[1 0]</code></td>
<td><code>[1 1]</code></td>
<td></td>
<td>( \gamma )</td>
</tr>
<tr>
<td><code>[1 0]</code></td>
<td><code>[0 0]</code></td>
<td></td>
<td>( \gamma )</td>
</tr>
<tr>
<td><code>[0 0]</code></td>
<td><code>[0 0]</code></td>
<td></td>
<td>( \gamma )</td>
</tr>
<tr>
<td><code>[0 0]</code></td>
<td><code>[0 0]</code></td>
<td></td>
<td>( \gamma )</td>
</tr>
</tbody>
</table>
Table 3. Best-reply graphs and NEs for all nine symmetric configurations of $U$.

<table>
<thead>
<tr>
<th>Defining Relations</th>
<th>Example $U$-shape</th>
<th>Best Replies</th>
<th>NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1(C,D) &lt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(0,0)$</td>
</tr>
<tr>
<td>$U_1(C,C) &lt; U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) = U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(p,0), (0,p)$</td>
</tr>
<tr>
<td>$U_1(C,C) &lt; U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) &gt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(0,1), (p,p), (1,0)$</td>
</tr>
<tr>
<td>$U_1(C,C) &lt; U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) &lt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(0,0), (1,1)$</td>
</tr>
<tr>
<td>$U_1(C,C) = U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) = U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(p,q)$</td>
</tr>
<tr>
<td>$U_1(C,C) = U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) &gt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(p,1), (1,p)$</td>
</tr>
<tr>
<td>$U_1(C,C) = U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) &lt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(0,0), (p,p), (1,1)$</td>
</tr>
<tr>
<td>$U_1(C,C) &gt; U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) = U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(0,0), (1,1)$</td>
</tr>
<tr>
<td>$U_1(C,C) &gt; U_1(D,C)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U_1(C,D) &gt; U_1(D,D)$</td>
<td></td>
<td></td>
<td>$(1,1)$</td>
</tr>
</tbody>
</table>

Note: Variables $p$ and $q$ represent all values in $[0, 1]$. 
Table 4. Symmetric 1-back SPEs for the IPD.

<table>
<thead>
<tr>
<th>U-shape</th>
<th>SPEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1(C, D) &lt; U_1(D, D)$</td>
<td>$\begin{bmatrix} 1 &amp; 0 \ 0 &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; 0 \ 0 &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; 0 \ 0 &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>$U_1(C, C) &gt; U_1(D, C)$</td>
<td>$\begin{bmatrix} p &amp; 0 \ 0 &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; 0 \ 0 &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; 0 \ 0 &amp; p \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; 0 \ 0 &amp; p \end{bmatrix}$</td>
</tr>
<tr>
<td>$U_1(C, D) &lt; U_1(D, D)$</td>
<td>$\begin{bmatrix} 1 &amp; p \ p &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; p \ p &amp; 0 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; p \ p &amp; 1 \end{bmatrix}$ vs. $\begin{bmatrix} 1 &amp; p \ p &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>$U_1(C, C) = U_1(D, C)$</td>
<td>$\begin{bmatrix} p &amp; p \ p &amp; p \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; p \ p &amp; p \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; p \ p &amp; p \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; p \ p &amp; p \end{bmatrix}$</td>
</tr>
<tr>
<td>$U_1(C, C) &lt; U_1(D, C)$</td>
<td>$\begin{bmatrix} p &amp; 0 \ 0 &amp; 1 \end{bmatrix}$ vs. $\begin{bmatrix} p &amp; 0 \ 0 &amp; 1 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

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 $\gamma$. 
Figure Captions.

**Figure 1.** Example best-reply diagram for a 2×2 game. Probability of choosing the focal action for player $k$ is denoted by $p_k$. The best reply for player 1 gives $p_1$ as a function of $p_2$ (dark line), and vice versa (light line). Intersections of the two lines (circled) are NEs.

**Figure 2.** Payoff shape corresponding to the best-reply graph in Figure 1. Circled points indicate expected payoffs associated with NEs. Dashed lines indicate the strategy of each player that makes the opponent’s payoff independent of his or her action.

**Figure 3.** Best-reply graph (A) and payoff shape (B) for the one-shot Prisoner’s Dilemma.

**Figure 4.** Example relationship between $U$ (heavy line) and $r$ (medium line) in the IPD. Dashed lines represent shrunken copies of the $U$-shape. Each small shape has its origin aligned with a vertex of the $r$-shape, and contains the corresponding vertex of the full-scale $U$-shape.

**Figure 5.** Relation of all-defect to other SPEs. Heavy lines indicate SPEs, with all-defect lying on the horizontal axis. In panel A, the solution \[
\begin{bmatrix}
 p & 0 \\
 0 & p
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
 p & 0 \\
 0 & p
\end{bmatrix}
\] only connects with all-defect via a discontinuous jump on the boundary of the parameter space. Here, the horizontal axis schematically represents the relationship between the boundary point $(x, y, \gamma) = (0, 1, 0)$ and the region of true IPDs. In panel B, the solution \[
\begin{bmatrix}
 0 & 0 \\
p & 0
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
 0 & p \\
0 & 0
\end{bmatrix}
\] tends continuously toward all-defect as $x$ approaches 0.
Figure 1
Figure 4

\[ U(C,C) \]
\[ U(D,D) \]
\[ U(C,D) \]
\[ U(D,C) \]
Figure 5

(A) \[
\begin{bmatrix}
  p & 0 \\
  0 & p
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
  p & 0 \\
  0 & p
\end{bmatrix}
\]

(B) \[
\begin{bmatrix}
  0 & 0 \\
  p & 0
\end{bmatrix}
\] vs. \[
\begin{bmatrix}
  0 & p \\
  0 & 0
\end{bmatrix}
\]