Evolutionary Stability in a Reputational Model of Bargaining

Dilip Abreu[*] Rajiv Sethi[†]

January 24, 2001

Abstract

A large and growing literature on reputation in games builds on the insight that the possibility of one or more players being other than fully rational can have significant effects on equilibrium behavior. This literature leaves unexplained the presence of behavioral players in the first place, and the particular forms of irrationality assumed. In this paper we endogenize departures from rationality on the basis of an evolutionary stability criterion, under the assumption that rational players incur a cost which reflects the greater sophistication of their behavior. This cost may be arbitrarily small. Within the context of a reputational model of bargaining, we show that evolutionary stability *necessitates* the presence of behavioral players, and places significant restrictions on the set of behavioral types that can coexist. It is consistent, however, with a broad variety of outcomes ranging from immediate agreement to complete surplus dissipation. The long run population share of behavioral types is greatest at states in which surplus dissipation is either negligible or almost complete.

[*] Department of Economics, Princeton University (dabreu@princeton.edu).

[†] Department of Economics, Barnard College, Columbia University (rs328@columbia.edu).

1 Introduction

A large and growing literature has emerged from the seminal contributions of Kreps and Wilson (1982) and Milgrom and Roberts (1982) on reputation formation in games. Equilibrium behavior in such models often accords more closely with the behavior of experimental subjects and is more intuitively appealing than behavior under common knowledge of rationality. For instance, cooperation can occur in most periods of a finitely repeated prisoners' dilemma (Kreps et al., 1982), entry can be deterred in early periods of the chain store game (Kreps and Wilson, 1982, Milgrom and Roberts, 1982), and efficient outcomes can be approximately attained in centipede games (McElvey and Palfrey, 1992). These results depend rather critically, however, on the nature of departures from rationality that are permitted within the model. For instance Kreps et al. (1982) consider a 'tit-for-tat' player in the iterated prisoners' dilemma and McElvey and Palfrey (1992) allow for an altruistic type who always plays 'across' in the centipede game. Alternative assumptions regarding the behavior of non-rational types may result in very different outcomes. For example, in the context of repeated games, Fudenberg and Maskin (1986) show that any feasible and individually rational payoff profile can be supported in equilibrium, by a suitable choice of behavioral type.

In this paper, we endogenize the nature and extent of departures from rationality on the basis of a process of evolutionary selection. We do so by interpreting the probability with which a player is of a particular type with the share of such types in a population from which the players are randomly drawn. The composition of this population determines the behavior of rational players and hence the expected payoffs of rational and behavioral players of each type. When there are payoff differences across types, the differential exerts evolutionary pressure on the population composition itself. Any long-run outcome of this process of evolutionary competition must be such that all surviving types have equal payoffs. Stable population states have the additional property that small perturbations do not result in cumulative divergence from the state.

When rational types can costlessly and perfectly mimic any behavioral type, they must obtain payoffs that are at least as great as the most profitable behavioral type. In general the payoffs of rational players will exceed those of the most profitable behavioral type at any nondegenerate population composition, in which case only rational players could survive in the long run. On the other hand, when the successful imitation of behavioral types is *not* perfectly costless, the long run population composition need not be degenerate. We allow for the possibility that rational players incur a cost that behavioral types avoid, and interpret this cost as arising from the greater flexibility, sophistication and information intensity of their behavior. In this latter case one may explore the nature of the population composition in evolutionary equilibrium. This is the central concern of the present paper.

Our analysis is conducted within the framework of a recent model of bargaining and reputation. Abreu and Gul (2000) have developed a reputation based theory of bargaining based on multiple behavioral types whom rational players may choose to imitate. Each behavioral type demands a particular share of the surplus to be divided and is unwilling to concede to any demand that is incompatible with this share. Equilibrium in this model is unique (with respect to outcomes) and is generally characterized by inefficient delays. The uniqueness of equilibrium is itself significant since any division of the surplus can be consistent with equilibrium when all players are rational.

A first and basic result is that the absence of behavioral types is incompatible with evolutionary stability. However small the cost of optimization that rational players incur, there does not exist an evolutionarily stable state in which behavioral types are entirely absent, whereas there do exist stable states in which rational types are not present. Furthermore, the criterion of evolutionary stability places significant restrictions on the set of behavioral types that can coexist. We show that if a behavioral type is present at an evolutionary equilibrium, then the type which makes the complementary demand must also be present. Thus an aggressive behavioral type (who demands more than half the surplus) must be counterbalanced by a correspondingly conciliatory type at any stable state. This in turn implies that it is impossible for all behavioral types to be aggressive.

Despite the fact that evolutionary stability significantly restricts the structure of population states, we show that it places almost no restrictions on the degree to which inefficient delays can occur. Stable states can be very close to efficient (with all behavioral types demanding close to half the surplus, and almost all interactions resulting in immediate agreement), or they may be very inefficient (with most behavioral types demanding almost the entire surplus and almost all interactions resulting in perpetual disagreement). At such extreme states rational players may be absent from the population even if the cost of optimization is very small. In addition, there typically exist stable states with intermediate levels of efficiency in which behavioral and rational players coexist. Hence evolutionary stability is consistent with a broad variety of outcomes ranging from no delays to an almost complete dissipation of the surplus, and with a significant presence or complete absence of rational players.

One interpretation of behavioral types is that they correspond to individuals who adhere to particular social conventions regardless of whether it is in their economic interest to do so. Under this interpretation, our results suggest that different societies facing the same objective conditions may end up not only with distinct conventions but also with significant differences in the degree to which their conventions are breached. There is a systematic and nonmonotonic relationship between the efficiency of the conventions that evolve and the share of the population that adheres to them. Specifically, rational behavior will be rarest in societies with conventions at either extreme of the efficiency scale, and most common in societies with modestly inefficient conventions.

We emphasize that in our analysis an individual's type is unobservable and can only be deduced from her actions along the equilibrium path. As is well known, players who exhibit *observable* departures from rationality can be successful under evolutionary competition with rational players because they can commit to actions that involve 'incredible' threats or promises (see, for instance, Banerjee and Weibull, 1995). There is no such possibility of commitment under complete unobservability, and behavioral types cannot therefore survive unless rationality comes at a cost. The evolutionary interaction of behavioral types with more sophisticated players under unobservability has previously been examined by Conlisk (1980), Dekel and Scotchmer (1992) and Stahl (1993). These papers differ in significant respects from the present work, most notably in the absence of reputational effects that induce rational players to imitate behavioral types. Guttman (1996) considers a reputational model of the finitely repeated prisoners' dilemma in which a 'tit-for-tat' type interacts with rational players, and the optimization cost has a variable component which increases in the length of the game. His work differs from ours in that his results entail optimization costs which are sufficiently large; furthermore since he assumes only a single reputational type the issue of what configurations of reputational types can survive and in what proportions, simply does not arise.

The paper is organized as follows. Section 2 describes the equilibrium behavior of rational players holding fixed the composition of the population from which players are drawn. Section 3 traces the resulting payoff implications for rational and behavioral players of each type, and introduces the criterion of evolutionary stability to be used in the remainder of the paper. The fact that evolutionary stability necessitates the presence of behavioral types is shown in Section 4, and additional restrictions on the set of behavioral types that can co-exist are obtained. Section 5 contains a complete characterization of stable states in the case of two behavioral types, and shows the range of demands, efficiency and rationality that can be consistent with evolutionary stability. Section 6 concludes.

2 Bargaining and Reputation in Equilibrium

Consider the following simple bargaining interaction. A surplus normalized to equal one unit is to be divided among two players. At time 0 the players simultaneously announce demands α_i and α_j. If $\alpha_i + \alpha_j \leq 1$, the demands are compatible and each player receives a share of the unit surplus that is proportional to her demand. If not, then either player may concede to the other's demand at any time $t \geq 0$ provided that neither player has already conceded. As soon as a concession occurs the game ends and each player receives the discounted value of their payoff at the time of concession. It is assumed that both players have the same rate of time preference r. Payoffs are determined as follows: if a single player concedes at time t to some demand α_i then her opponent's discounted payoff is $\alpha_i e^{-rt}$ while her own payoff is $(1 - \alpha_i)e^{-rt}$. If both players concede simultaneously at time t, the combined payoff is e^{-rt} and their individual payoffs are in proportion to their demands.

Players may be either *rational* or *behavioral.* At any point in time there is a finite set of behavioral types denoted $C = \{\alpha_1, ..., \alpha_n\}$. A behavioral player of type i always demands a share $\alpha_i \in C$ of the surplus, accepts any offer greater than or equal to this demand, and (perpetually) rejects smaller offers. The probability that any given player is behavioral of type i is denoted $z_i > 0$, and $z_0 = 1 - \sum_{k=1}^{n} z_k > 0$ is the probability that the player is rational. Note that these probabilities are the same for both players. We interpret the probabilities as population shares of each type in a large population from which the two players are randomly drawn, and refer to the vector $z = (z_1, ..., z_n)$ as the *population composition.* This is an element of the set $\Omega^n = \{z \in R^{n+1} \mid z_i > 0 \text{ and } \sum_{i=1}^{n} z_i < 1\}$. We shall refer to the pair (C, z), which specifies the demands made by behavioral types as well as their shares in the population as the *population state.*

A strategy for player 1 consists of a probability distribution μ^1 on C and a set of cumulative distributions $F_{ij}^1(t)$ on $R_+ \cup \{\infty\}$ for all pairs i, j such that $\alpha_i + \alpha_j > 1$. Here $F_{ij}^1(t)$ is the probability of player 1 conceding to her opponent's demand by time t, given that the two players have chosen the incompatible demands α_i and α_j respectively. Similarly, a strategy for player 2 is given by a probability distribution μ^2 on C and a set of cumulative distributions $F_{ji}^2(t)$ on $R_+ \cup \{\infty\}$ for all pairs i, j such that $\alpha_i + \alpha_j > 1$. This describes a symmetric version of the continuous time bargaining game examined by Abreu and Gul (2000), who show that equilibrium is unique. Hence the equilibrium expected payoffs of all (rational and behavioral) players are determined uniquely as a function of the (commonly known) population composition z. Given the symmetry assumed here (which itself is a consequence of assuming that the two players are randomly drawn from the *same* population), we drop superscripts and let μ_i denote the probability with which α_i is chosen by a rational player in equilibrium. Similarly, $F_{ij}(t)$ denotes the probability that a player demanding α_i concedes to her opponent's incompatible demand α_j by time t.

Abreu and Gul (2000) establish that equilibrium in this game has the following structure. The probability that a player is behavioral conditional on the fact that she chooses demand α_i is given by

$$p_i = \frac{z_i}{z_i + z_0\mu_i}$$

If the demands made by the two players are incompatible, then at most one player concedes to the other's demand with positive probability at time 0. That is, if $\alpha_i + \alpha_j > 1$ then either $q_{ij} = 0$ or $q_{ji} = 0$ (or both), where $q_{ij} = F_{ij}(0)$ is the probability of immediate concession by a player demanding α_i to a player demanding α_j. If concession does not occur at time zero, a war of attrition results.[1] A player who has made demand α_i and faces an opponent who has made demand α_j, concedes at a constant hazard rate λ_{ij}, where

$$\lambda_{ij} = \frac{r\left(1-\alpha_i\right)}{\alpha_j - \left(1-\alpha_i\right)}. \tag{1}$$

If neither player concedes immediately to the other's demand with positive probability ($q_{ij} = q_{ji} = 0$), then the probability that the player choosing α_i is behavioral conditional on the fact that she has not yet conceded increases over time and reaches 1 at time T_{ij},where

$$T_{ij} = -\frac{\log p_i}{\lambda_{ij}}. \tag{2}$$

If $T_{ij} = T_{ji}$ (the probability a player is behavioral reaches 1 simultaneously for both players) then $q_{ij} = q_{ji} = 0$ in equilibrium. If not, then one of the players concedes immediately to the other's demand with positive probability. This probability of immediate concession must be precisely such as to cause the probability that each player is behavioral to reach 1 simultaneously for both players. This occurs at time $T_{ij}^0 = \min\{T_{ij}, T_{ji}\}$. Hence, if $T_{ij} > T_{ji}$ then $q_{ij} > 0 = q_{ji}$ (the player demanding α_i concedes immediately with positive probability to the player demanding α_j).

A number of results characterizing additional properties of equilibrium behavior in this bargaining game are presented in the appendix, including the following: (i) if a particular behavioral type is imitated by rational players then all behavioral types making more aggressive demands are also imitated, and (ii) Any type which is incompatible with all types

[1] Rather than being assumed, this war of attrition structure is *derived* by Abreu and Gul as a property of equilibrium in the the limit of a sequence of dicrete time bargaining games.

(including itself) must be imitated by rational players and will never be immediately conceded to with positive probability. An immediate corollary of the latter is that if all types are aggressive (in the sense of demanding more than half the surplus), then all will be imitated by rational players with probabilities that imply no immediate concessions in equilibrium. These properties are implicit in the work of Abreu and Gul (2000) and, while tangential to the present inquiry, are required in proving our main results on evolutionary stability.

3 Payoffs

Since the population state (C, z) uniquely determines the equilibrium behavior of rational players, it also uniquely determines their expected payoffs and those of behavioral types. Since we wish to endogenize the determination of the population state, payoff functions need to be defined in a manner that is general enough to encompass different states (and not just different population compositions for a given state). Accordingly, let $\mathcal{C}$ denote the set of finite subsets of $(0, 1)$ and let $\mathcal{W} = \{\emptyset, \Omega^1, \Omega^2, ...\}$ denote the set of finite dimensional unit simplexes. Define the set $\mathcal{S}$ as follows:

$$\mathcal{S} = \{(C, \Omega^n) \in \mathcal{C} \times \mathcal{W} \mid n = |C|\}.$$

All possible population states are elements of $\mathcal{S}$, and each element of $\mathcal{S}$ corresponds to a population state. The payoffs of the different types present at a population state may be defined by the functions $\pi_i : \mathcal{S} \rightarrow R$, for $i = 0, 1, ..., n$. Here $\pi_0(C, z)$ denotes the expected payoff to rational players at state (C, z) and, for $i \geq 1$, $\pi_i(C, z)$ denotes the expected payoff to a behavioral player who demands α_i at population state (C, z).[2]

We shall allow for the possibility that rational players, on account of the greater sophistication of their behavior, incur a cost $\varepsilon > 0$ which is deducted from their payoffs from the bargaining game. The explicit inclusion of such a cost reflects a rudimentary attempt to correct for the fact that models of unbounded rationality neglect an important scarce resource, the computational capacity of the human mind (Simon, 1978). Our qualitative results do not depend on the magnitude of this cost: in particular, ε may be taken to be arbitrarily small.

[2] Note that $\mathcal{S}$ includes the state in which C and z are both empty (only the rational type is present in the population). At this state equilibrium behavior is not uniquely determined and so the payoff functions are not well defined. As shown below, this does not prevent us from investigating the stability of this state since payoffs are well defined at all other states in any neighborhood of this state.

We wish to identify population states that are the long run outcome of a process of evolutionary selection. We do this by employing the notion of an evolutionarily stable state (Maynard Smith and Price, 1973). In order for a state to be stable in this sense, it must be able to resist small perturbations in the population state. These perturbations may be of two types: (i) changes in the population composition, holding fixed the set of demands C, and (ii) changes in the set of demands itself.

A first step in establishing criteria for evolutionary stability is to identify, for a given set of demands C, states that are stable with respect to small perturbations in the population composition z. Given some set of demands C, consider any two population compositions z and z', and define $f(z, z')$ as follows

$$f(z, z') = \sum_{i=1}^{n} z_i \pi_i(C, z').$$

This may be interpreted as the expected payoff to a player drawn from a population with composition z when the opponent she faces is drawn from a population with composition z' (with both populations having the same set of possible behavioral demands C). We say that the population state (C^*, z^*) is *stable with respect to perturbations in the population composition* if, for every $z' \in \Omega^n$ with $z' \neq z^*$, there exists some $\bar{\eta} > 0$ such that

$$f(z^*, (1-\eta)\, z^* + \eta z') > f(z', (1-\eta)\, z^* + \eta z')$$

for all $\eta \in (0, \bar{\eta})$. The interpretation is that a population with composition z^* will obtain a greater expected payoff than one with any other composition z' against a mixture of the two in which the latter has sufficiently small weight. The main justification for the use of this condition is that any population state z^* which satisfies it is asymptotically stable under a variety of evolutionary selection dynamics including the replicator dynamics. An alternative version of the condition is that there exists some neighborhood $N(z^*) \subset \Omega^n$ of z^* such that for all $z \in N(z^*)$ with $z \neq z^*$, the following holds:

$$f(z^*, z) > f(z, z). \tag{3}$$

This requires that the "strategy" z^* obtain a strictly greater payoff against all nearby strategies z than these strategies get when matched against themselves. This latter version of the condition will be used in the analysis below.[3]

[3] See Hofbauer and Sigmund (1988) for the equivalence of the two definitions, and the relationship between this static notion of stability and dynamic models of evolutionary selection.

While the above criterion identifies population states that are stable with respect to small perturbations in the population composition for a given set of behavioral demands C, it does not address the question of whether such a state would be stable with respect to perturbations in the set of demands itself. It is amply possible to find population states that satisfy the condition when the set of possible demands C is restricted, but which would be unstable in the presence of a richer set of behavioral types.[4] To address this problem, consider any population state $(C^*, z^*) \in \mathcal{S}$ in which the number of behavioral types is n. Let $\alpha_{n+1} \in (0, 1)$ where $\alpha_{n+1} \notin C^*$ represent the demand of some 'mutant' behavioral type and let $C = C^* \cup \{\alpha_{n+1}\}$ denote the enlarged set of types. Let $w = ((1-\eta)z^*, \eta) \in \Omega^{n+1}$ be a state with weight $\eta > 0$ on the population share of the "mutant" type, and weight $(1-\eta)$ on the incumbent population z^*. The payoffs to each type (including the mutant type) are uniquely determined at any such population state w. We shall say that the state (C^*, z^*) is *uninvadable* if, for every $\alpha_{n+1} \notin C^*$, there exists some $\bar{\eta} > 0$ such that, for all $\eta < \bar{\eta}$ and all $i \in \{1, ..., n\}$

$$\pi_i(C, w) > \pi_{n+1}(C, w). \tag{4}$$

In other words, a state (C^*, z^*) is uninvadable if all potential mutants earn a strictly lower payoff than the incumbents when the population state is sufficiently close to (C^*, z^*).

Any state which is stable with respect to perturbations in the population composition and is also uninvadable is said to be *evolutionarily stable*.[5] Any state which is not evolutionarily stable is *unstable*. We turn next to the question of which population states are consistent with evolutionary stability.

4 Evolutionary Stability

To build intuition for the results to be presented below, consider first a population in which rational players coexist with a single behavioral type who demands $x > \frac{1}{2}$ and has population share $\eta \in (0, 1)$. The population state is therefore (C, z) where $C = \{x\}$ and $z = \eta$. At this state rational players choose x with probability 1 since a failure to do so would reveal

[4] For instance, if C consists of a single demand $\alpha_1 > \frac{1}{2}$, there always exists a state which is stable with respect to perturbations in the population composition but which is no longer stable if the set of demands is suitably enlarged. This example is discussed further below.

[5] The criterion of evolutionary stability is typically applied with some exogenously given *finite* set of behaviors, in which case the uninvadability condition can be absorbed into the stability condition (3). We separate the two in order to allow for types making any demand in the unit interval.

their rationality and force them to concede immediately to their opponents demand. Since all players demand x, the demand conveys no information about a player's type and posterior probability that a player is behavioral immediately following the demand is simply η. There are no immediate concessions and a war of attrition follows. From (1) and (2), each player concedes at a hazard rate $\lambda = r\,(1-x)\,/\,(2x-1)$ and the probability that a player is behavioral conditional on this player not having conceded reaches 1 at time $T = (-\log\eta)/\lambda$.

The expected payoff to rational players from this bargaining interaction is $1-x$, since immediate concession is in the support of the equilibrium strategy. Concession with probability 1 at time T (and not earlier) is also in the support of the equilibrium strategy, so this must also yield an expected payoff of $1-x$. We therefore have

$$1-x = (1-\eta)\,\theta + \eta\,(1-x)\,e^{-rT},$$

where θ denotes the payoff that concession with probability 1 at time T yields when one's opponent is rational (and must therefore have conceded with probability 1 by the time T is reached) and $(1-x)\,e^{-rT}$ is the payoff that concession with probability 1 at time T yields when one's opponent is behavioral (and therefore never concedes). Clearly, θ is also the payoff that a behavioral player would get when matched with a rational opponent. Hence the expected payoff to behavioral types is

$$\pi_1(C,z) = (1-\eta)\,\theta = (1-x)\left(1-\eta e^{-rT}\right), \tag{5}$$

The term $(1-x)\,\eta e^{-rT}$ represents the difference between the payoffs of rational and behavioral types from the bargaining interaction. This term can be interpreted as the 'penalty' that behavioral types pay because, unlike rational players, they fail to concede even when their opponent is revealed to be behavioral. *This penalty approaches* 0 *as the population share of behavioral types vanishes*: $\lim_{\eta\to 0}\pi_1(C,z) = 1-x$. Since $\pi_0(C,z) = 1-x-\varepsilon$, there exists $\bar{\eta} > 0$ such that $\pi_0 < \pi_1$ for all $\eta < \bar{\eta}$. Hence the behavioral type obtains a strictly greater payoff than the rational type when the population share of the former is sufficiently small. For any $\varepsilon > 0$, the state in which only rational types are present can be invaded.

The reason why a population of rational players cannot expel a rare aggressive mutant is the following. The behavioral strategy of never conceding is only costly when behavioral types encounter each other. The fact that behavioral players fail to concede even when it becomes certain that their opponent is behavioral is immaterial when they encounter rational players since this point in time is never reached. As the population share of behavioral types falls, it becomes increasingly unlikely that behavioral types encounter each other. In the limit, the payoffs of behavioral and rational player from the bargaining interaction converge, so that even an arbitrarily small cost of rationality can give the former an advantage.

Since a population of rational players is unstable in the presence of a single aggressive behavioral type, this raises the question of whether a population consisting of such a behavioral type together with the rational type can itself be stable. It turns out that it cannot. To see why, consider a population state (C, z) in which there are two behavioral types, both of whom demand more than half the surplus. We may write $C = \{\alpha_1, \alpha_2\}$ and suppose, without loss of generality, that $\alpha_2 > \alpha_1 > \frac{1}{2}$. The population composition is $z = (z_1, z_2)$. At an equilibrium, both types are imitated with positive probability and there are no immediate concessions with positive probability in equilibrium.[6] Since immediate concession is in the support of the a rational player's equilibrium strategy regardless of her opponent's demand, the expected payoffs to rational types are

$$\pi_0(C, z) = (z_1 + z_0\mu_1)(1 - \alpha_1) + (z_2 + z_0\mu_2)(1 - \alpha_2).$$

Let T_{ij} denote the time at which a player demanding $\alpha_i \in C$ is revealed to be behavioral when her opponent has demanded $\alpha_j \in C$. Since there are no immediate concessions, $T_{ij} = T_{ji}$. Using the same reasoning as was used to obtain (5), we obtain

$$\begin{aligned}
\pi_1(C, z) &= (z_1 + z_0\mu_1)(1 - \alpha_1)\left(1 - p_1 e^{-rT_{11}}\right) + (z_2 + z_0\mu_2)(1 - \alpha_2)\left(1 - p_2 e^{-rT_{12}}\right) \\
\pi_2(C, z) &= (z_1 + z_0\mu_1)(1 - \alpha_1)\left(1 - p_1 e^{-rT_{21}}\right) + (z_2 + z_0\mu_2)(1 - \alpha_2)\left(1 - p_2 e^{-rT_{22}}\right)
\end{aligned}$$

where p_i is the posterior probability that player is behavioral at time 0 conditional on her having demanded α_i. From (1) and the fact that $\alpha_2 > \alpha_1$, we have $\lambda_{2j} < \lambda_{1j}$ for each $j \in \{1, 2\}$. This in turn implies from (2) that $T_{2j} > T_{1j}$ for each $j \in \{1, 2\}$ and hence, from the above equations, that $\pi_2(C, z) > \pi_1(C, z)$. In other words, the more aggressive type obtains a greater payoff than the less aggressive type at any population composition in which both are present. This payoff advantage arises because, in the case of the more aggressive type, the 'penalty' for not conceding when one's opponent is revealed to be behavioral is paid *further into the future.* Clearly no population state with a single aggressive behavioral type can therefore be evolutionarily stable, since a type with a more aggressive demand can invade.

The above argument can be generalized to show that no population in which the behavioral types all demand more than half the surplus can be stable, because a type that is even more aggressive can invade. Consider, however, a state in which *not* all types are aggressive, and in which the most aggressive and the least aggressive types make complementary

[6] This follows from Lemma 2 in the Appendix.

demands. In this case, if an even more aggressive 'mutant' were to enter the population, it is not necessarily the case that the latter would earn a higher payoff than the incumbents. The reason is that the payoff advantage identified above is counteracted by the fact that there the mutant demand, unlike that of the incumbents, is incompatible with that of the least aggressive type. This suggests that the behavioral types present at any evolutionarily stable state must consist of complementary pairs. This is indeed the case.

Theorem 1 *A population state (C^*, z^*) is evolutionarily stable only if, for each demand $\alpha_i \in C^*$ there exists a complementary demand $\alpha_j = 1 - \alpha_i \in C^*$.*

Proof. See appendix.

An immediate consequence of Theorem 1 is that at any stable state, there must be at least one behavioral type which demands at most half the surplus.

While evolutionary stability restricts the set of possible behavioral types that can coexist in significant and interesting ways, it allows for a very broad range of outcomes with regard to surplus dissipation and welfare. We show in the next section that states with an almost complete dissipation of the surplus and no rational types present can also be stable, as can states which have a significant presence of rational types and are more modestly inefficient. We do this by characterizing the set of stable states that contain exactly two behavioral types.

5 Two Behavioral Types

Consider a population with two complementary behavioral types, that is, $C = \{\alpha_1, \alpha_2\}$ where $\alpha_1 = x \in (\frac{1}{2}, 1)$ and $\alpha_2 = 1 - x$. Suppose further that the population composition z is such that $z_1 + z_2 = 1$, so no rational players are present. The payoffs to the two behavioral types are, respectively,

$$\begin{aligned} \pi_1 &= xz_2 \\ \pi_2 &= (1-x)(1-z_2) + \frac{1}{2}z_2 \end{aligned}$$

At any state satisfying (3), equality of payoffs implies $z_2^* = 2 - 2x$ and $z_1^* = 2x - 1$. The (common) payoff to each type is $(2-2x)x$. Let $s^* = (C, z^*)$ and observe that for any $s = (C, z)$, we have $f(s^*, s^*) = f(s, s^*)$ and

$$f(s^*, s) - f(s, s) = (z_1^* - z_1)xz_2 + (z_2^* - z_2)\left((1-x)(1-z_2) + \frac{1}{2}z_2\right).$$

Since $(z_1^* - z_1) = -(z_2^* - z_2)$ we have

$$\begin{aligned} f(s^*, s) - f(s, s) &= (z_2^* - z_2)\left((1-x)(1-z_2) + \frac{1}{2}z_2 - xz_2\right) \\ &= (2 - 2x - z_2)\left((1-x)(1-z_2) + \frac{1}{2}z_2 - xz_2\right) \\ &= \frac{1}{2}(2 - 2x - z_2)^2 \geq 0 \end{aligned}$$

with strict equality holding when $z_2 \neq 2 - 2x = z_2^*$. Hence z^* satisfies (3).

To see that z^* is uninvadable, consider a population state $w = ((1-\eta)z^*, \eta)$ in which a behavioral 'mutant' type demanding $\alpha_3 = y$ is present with population share $\eta \in (0,1)$, and let $C' = \{\alpha_1, \alpha_2, \alpha_3\}$. Payoffs to the mutant are as follows:

$$\pi_3(C', w) = \begin{cases} 0 & \text{if } y > x \\ \left(\frac{y}{1-x+y}\right) z_2 & \text{if } y \in (\frac{1}{2}, x) \\ \left(\frac{y}{1-x+y}\right) z_2 + \frac{1}{2}\eta & \text{if } y \in (1-x, \frac{1}{2}] \\ \left(\frac{y}{x+y}\right) z_1 + \left(\frac{y}{1-x+y}\right) z_2 + \frac{1}{2}\eta & \text{if } y \in (0, 1-x) \end{cases}$$

In the first two cases $\pi_3 < xz_2 = \pi_1$ for all $\eta \in (0,1)$. In the third case, if η is sufficiently small, $\pi_3 < xz_2 = \pi_1$. In the fourth case $\pi_3 < (1-x)z_1 + \frac{1}{2}z_2 + \left(\frac{1-x}{1-x+y}\right)\eta = \pi_2$. In each case, there exists $\bar{\eta} > 0$ such that for all $\eta < \bar{\eta}$ the difference between π_3 and $\max\{\pi_1, \pi_2\}$ is bounded away from 0. Since $\lim_{\eta \to 0}(\pi_1 - \pi_2) = 0$, there exists $\bar{\eta} > 0$ such that for all $\eta < \bar{\eta}$, we have $\pi_3 < \min\{\pi_1, \pi_2\}$. Hence z^* is uninvadable by behavioral types.

To show that z^* is uninvadable by the rational type, consider a population composition $z = (1-\eta)z^*$, in which the population share of rational types is η. It is possible to show (see Lemma 4 in the appendix) that rational players will never imitate the less aggressive behavioral type. The payoff to a rational player therefore satisfies

$$\begin{aligned} \lim_{\eta \to 0} \pi_0(C, z) &= \lim_{\eta \to 0}((1-x)(1-(1-\eta)z_2^*) + (1-\eta)xz_2^* - \varepsilon) \\ &= (1-x)(1-(2-2x)) + x(2-2x) - \varepsilon \end{aligned}$$

while $\lim_{\eta \to 0} \pi_1 = \lim_{\eta \to 0} \pi_2 = (2-2x)x$. If

$$\varepsilon > (1-x)(2x-1) \tag{6}$$

there exists $\bar{\eta} > 0$ such that rational players obtain lower payoffs that the two behavioral types for all $\eta < \bar{\eta}$. Rational players therefore cannot invade. On the other hand, if (6) is reversed, rational players can invade and z^* will not be evolutionarily stable. Since $x \in (0,1)$, inequality (6) is necessarily satisfied if $\varepsilon > \frac{1}{8}$. If, on the other hand, $\varepsilon \in (0, \frac{1}{8})$, the range of values for which (6) is violated comprises an interval:

$$(x_l, x_h) = \left(\frac{3}{4} + \frac{1}{4}\sqrt{(1-8\varepsilon)}, \frac{3}{4} - \frac{1}{4}\sqrt{(1-8\varepsilon)}\right)$$

Note that $x_l > \frac{1}{2}$ and $x_h < 1$ for all $\varepsilon \in (0, \frac{1}{8})$, and that the interval "shrinks" (x_l rises and x_h falls) as ε rises. We have therefore proved that any rest point with two complementary behavioral types demanding $x \in (\frac{1}{2}, 1)$ and $1-x$ respectively is evolutionarily stable provided that $x \in (\frac{1}{2}, x_l)$ or $x \in (x_h, 1)$.[7] In other words, if the aggressive type is sufficiently egalitarian or sufficiently aggressive, a population consisting of two complementary behavioral types and no rational types is evolutionarily stable. This is a special case of the following.

Theorem 2 *For every $x \in (\frac{1}{2}, 1)$, there exists an evolutionarily stable state with exactly two complementary behavioral types who demand x and $1-x$ respectively. At any such state rational types will be present if and only if $x \in (x_l, x_h)$.*

Proof. See appendix.

Theorem 2 states that *any* aggressive behavioral type may be present at an evolutionarily stable state provided that its complementary type is also present. If the type makes a demand that is sufficiently close to half the surplus, the corresponding state will resist invasion by *rational* types. The same is true if the demand is sufficiently close to the entire surplus. The existence of stable states without rational players occurs regardless of the magnitude of the optimization cost, although the range of demands which can be found at such states depends on this cost.

[7] It is easily seen that the limit case of a population consisting only of a single behavioral type which demands exactly half the surplus is also evolutionarily stable. The rational type cannot invade because, in any population with rational players and the incumbent behavioral type, the former will imitate the single behavioral type with certainty (since a failure to do so would result in the player being revealed to be rational). Each type gets $\frac{1}{2}$ in each interaction and with any $\varepsilon > 0$ the rational type obtains a strictly lower payoff than the incumbent. To see that other behavioral types cannot invade, consider some mutant demanding $\alpha_2 \neq \frac{1}{2}$. If $\alpha_2 > \frac{1}{2}$ the mutant obtains 0 in all interactions which is strictly less than the payoff of the incumbent. If $\alpha_2 < \frac{1}{2}$ the mutant obtains strictly below $\frac{1}{2}$ in all interactions while the incumbent obtains at least $\frac{1}{2}$. Hence the population is evolutionarily stable.

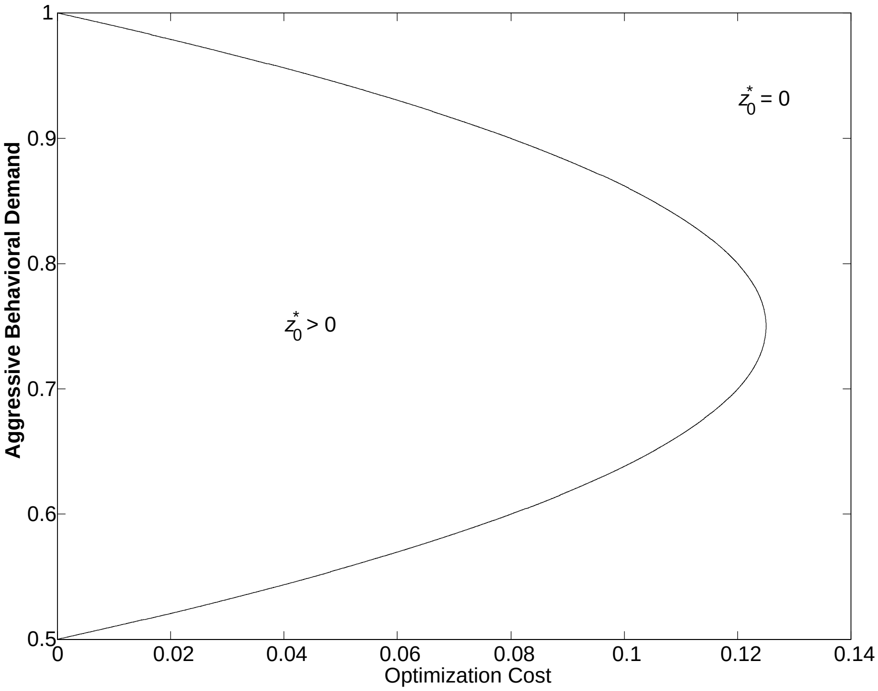

Figure 1. Range of demands for which rational types survive

Figure 1 illustrates the manner in which the interval (x_l, x_h) changes with the optimization cost ε. At stable states in which the aggressive behavioral demand lies within the upper and lower bounds, and only at such states, will rational types be present in the population. Higher values of ε correspond to a smaller range of demands for which rational players can survive, but even for ε close to zero, there exists a range of demands for which rational players do not survive. The special case of $\varepsilon = 0.04$ is depicted in Figure 2. Here $(x_l, x_h) \approx (0.54, 0.96)$. Corresponding to any aggressive demand within this range, there exists a stable state at which behavioral and rational types coexist. Corresponding to any aggressive demand outside this range, there exists a stable state at which only behavioral types survive. The value of ε places an upper bound (but clearly no lower bound) on the population share of rational players that is consistent with evolutionary stability.

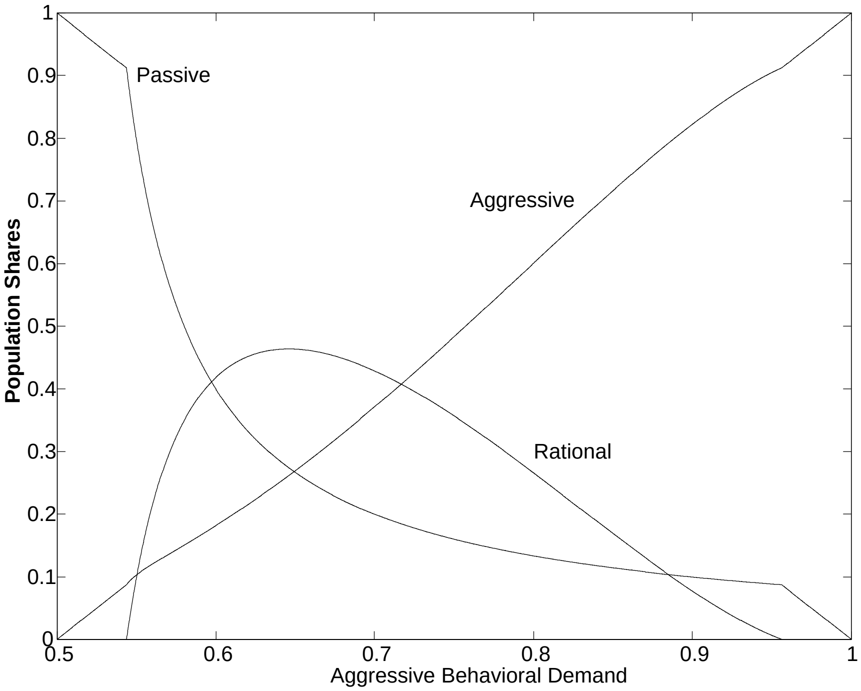

Figure 2. Population shares of the three types ($\varepsilon = 0.04$).

If one views behavioral types as individuals who adhere to social conventions even when it is not in their economic interest to do so, one could interpret Theorem 2 as asserting the existence of a range of stable but distinct societies which may be Pareto-ranked in their levels of efficiency and welfare. The incidence of conventional (as opposed to rational) behavior will be greatest in those societies with conventions that are either very efficient or very inefficient. Societies that stabilize at moderately inefficient conventions will also be those in which rational types are most prevalent, and hence also those in which conventions are most likely to be breached.

6 Conclusions

The analysis in this paper may be seen as an attempt to explore the implications of evolutionary competition among different behaviors when rationality is costly and reputational effects are important. In the bargaining environment examined here, we find that an endogenous explanation for the presence of behavioral types does not require that the cost of rationality be significant. No matter how small such costs happen to be, not only must behavioral types be present at stable states, but there must exist stable states in which rational types are absent. While evolutionary stability substantially restricts the range of behavioral types that can *co-exist,* it allows for the stability of any behavioral type provided that it's complementary type is also present. A wide range of demands and delays are therefore consistent with evolutionary equilibrium.

The basic methodology adopted here could, in principle, be applied to the endogenization of behavior in any reputational model with behavioral types. To do so, however, requires a reasonable specification of all possible behavioral types which could potentially survive the process of evolutionary competition. One could then ask whether the particular departures from rationality that have been assumed in the literature can be justified on evolutionary grounds.

Appendix

Prior to proving Theorems 1 and 2, we present the following results which characterize equilibrium behavior in the bargaining game for a given population state (C, z). The first two results are implicit in the work of Abreu and Gul (2000).

Lemma 1 *If* $\mu_i > 0$ then $\mu_j > 0$ for all $\alpha_j > \alpha_i$.

Proof. Suppose $\mu_i > 0$ and $\mu_j = 0$ for some $\alpha_j > \alpha_i$. Then a rational player who demands α_j with certainty and concedes to all demands time $\tau > 0$ will obtain at least

$$\sum_{\alpha_k \geq 1-\alpha_j} \left(z_0 \mu_k \alpha_j + z_k e^{-r\tau} \left(1-\alpha_k\right)\right) + \sum_{\alpha_k < 1-\alpha_j} \left(z_0 \mu_k + z_k\right) \frac{\alpha_j}{\alpha_j + \alpha_k}$$

The maximum possible payoff that rational players imitating α_i can obtain is

$$\sum_{\alpha_k \geq 1-\alpha_i} \left(z_0 \mu_k \alpha_i + z_k \left(1-\alpha_k\right)\right) + \sum_{\alpha_k < 1-\alpha_i} \left(z_0 \mu_k + z_k\right) \frac{\alpha_i}{\alpha_i + \alpha_k}$$

For τ sufficiently small, the deviation is strictly profitable, contradicting the hypothesis that $\mu_j = 0$ in equilibrium. Hence $\mu_j > 0$. ■

Lemma 2 *Suppose* $\alpha_i \in C$ *and* $\alpha_i + \alpha_k \geq 1$ for all $\alpha_k \in C$. *Then* $\mu_i > 0$ *and* $q_{ki} = 0$ *for all* $\alpha_k \in C$.

Proof. Suppose $\alpha_i \in C$ with $\alpha_i + \alpha_k \geq 1$ for all $\alpha_k \in C$ and $\mu_i = 0$. Consider a rational player who switches to the following strategy: make demand α_i with probability 1 and concede with probability 1 at time $\tau > 0$. This player will obtain a payoff $\alpha_i > \frac{1}{2}$ against all rational opponents since she will be believed to be behavioral with certainty, and a payoff of at least $(1-\alpha_k)\, e^{-r\tau}$ against behavioral players. Since the expected payoff of rational players against each other is at most $\frac{1}{2}$, and the expected payoff of rational players against behavioral players of type k is at most $(1-\alpha_k)$, this deviation will be strictly profitable if τ is sufficiently small. Hence $\mu_i = 0$ is inconsistent with equilibrium and we must have $\mu_i > 0$.

From (1) and (2), observe that

$$T_{jk} \geq T_{kj} \qquad \Leftrightarrow \qquad p_j^{\frac{1}{1-\alpha_j}} \leq p_k^{\frac{1}{1-\alpha_k}}. \tag{7}$$

Hence there must be at least one type j such that $T_{jk} \geq T_{kj}$ for all other types k. Such a player must be imitated by rational types (otherwise $T_{jk} = 0 < T_{kj}$ for any player k that is imitated by rational types), and is never conceded to with positive probability at time 0. Hence rational players imitating this type obtain an expected payoff of *at most* $1-\alpha_k$ against opponents who demand α_k. A rational player making demand α_i gets a payoff of *at least* $1-\alpha_k$ against players demanding k. Since types i and j are both imitated in equilibrium, a rational player making demand α_i must get exactly $1-\alpha_k$ against players demanding k. This can only occur if $q_{ki}=0$ for all k. ■

Lemma 3 *For any α_i, $\alpha_j \in C$, if $q_{ij} \geq q_{ji}$ then $q_{kj} \geq q_{ki}$ for all $\alpha_k \in C$.*

Proof. Suppose $q_{ij} \geq q_{ji}$. Then $T_{ij} \geq T_{ji}$ and from (7)

$$p_i^{\frac{1}{1-\alpha_i}} \leq p_j^{\frac{1}{1-\alpha_j}}. \tag{8}$$

Hence, for any $\alpha_k \in C$ either (i) $q_{ki}=0=q_{kj}$ or (ii) $q_{ki}=0<q_{kj}$ or (iii) $q_{ki}>0$ and $q_{kj}>0$. For the first two cases the result is immediate, so suppose $q_{ki}>0$ and $q_{kj}>0$. Then $T_{ik}=\tilde{T}_{ki}<T_{ki}$ where

$$\tilde{T}_{ki}=-\frac{\log \tilde{p}_{ki}}{\lambda_{ki}}$$

and

$$\tilde{p}_{ki}=\frac{z_k}{z_k+z_0\mu_k\left(1-q_{ki}\right)}.$$

Here $\tilde{p}_{ki}$ is the posterior probability with which a player is behavioral conditional on their having demanded α_k and having failed to concede immediately to demand α_i. Hence

$$\left(\frac{z_k}{z_k+z_0\mu_k\left(1-q_{ki}\right)}\right)^{\frac{1}{\lambda_{ki}}}=p_i^{\frac{1}{\lambda_{ik}}}$$

or

$$\left(\frac{z_k}{z_k+z_0\mu_k\left(1-q_{ki}\right)}\right)^{\frac{1}{1-\alpha_k}}=p_i^{\frac{1}{1-\alpha_i}}$$

Similarly, since $q_{kj}>0$,

$$\left(\frac{z_k}{z_k+z_0\mu_k\left(1-q_{kj}\right)}\right)^{\frac{1}{1-\alpha_k}}=p_j^{\frac{1}{1-\alpha_j}}.$$

From the previous two equations and (8), it follows that $q_{kj} \geq q_{ki}$. ■

Lemma 4 Suppose $C=\{\alpha_1,\alpha_2,\alpha_3\}$ with $\frac{1}{2}<\alpha_1<\alpha_3$ and $\alpha_2=1-\alpha_1$, and $(z_1,z_2,z_3)=((1-\eta)\,z^*,\eta)$ where $\eta>0$ and $z^* \in \Omega^2$. Then there exists $\bar{\eta}>0$ such that, for all $\eta<\bar{\eta}$, $\mu_2=0$.

Proof. Let $\alpha_1 = x$ and $\alpha_2 = y > x$. First we show that $\lim_{\eta \to 0} \mu_3 = 0$. Suppose, by way of contradiction, that $\lim_{\eta \to 0} \mu_3 > 0$. Then we would have $\lim_{\eta \to 0} p_3 = 0$ and $\lim_{\eta \to 0} T_{31} = \lim_{\eta \to 0} T_{32} = \infty$. This implies that there exists $\bar{\eta} > 0$ such that for all $\eta < \bar{\eta}$ we have $q_{31} > 0$ and $q_{32} > 0$. In this case the payoffs to rational players demanding x and rational players demanding y are, respectively,

$$\begin{aligned} \rho_1 &= (1-x)(z_1 + z_0\mu_1) + x(z_2 + z_0\mu_2) + \rho_{13}(z_3 + z_0\mu_3), \\ \rho_3 &= (1-x)(z_1 + z_0\mu_1) + x(z_2 + z_0\mu_2) + (1-y)(z_3 + z_0\mu_3) \end{aligned}$$

where $\rho_{13} > 1 - y$ and hence $\rho_1 > \rho_3$. But this implies $\mu_3 = 0$ for all $\eta < \bar{\eta}$, contradicting $\lim_{\eta \to 0} \mu_3 > 0$. Hence $\lim_{\eta \to 0} \mu_3 = 0$.

The payoffs to rational players demanding x and rational players demanding $1 - x$ are, respectively,

$$\begin{aligned} \rho_1 &= (1-x)(z_1 + z_0\mu_1) + x(z_2 + z_0\mu_2) + \rho_{13}(z_3 + z_0\mu_3), \\ \rho_2 &= (1-x)(z_1 + z_0\mu_1) + \frac{1}{2}(z_2 + z_0\mu_2) + \rho_{23}(z_3 + z_0\mu_3). \end{aligned}$$

Since $\lim_{\eta \to 0} \mu_3 = 0$, we have $\lim_{\eta \to 0}(z_3 + z_0\mu_3) = 0$. Since $\lim_{\eta \to 0}(z_2 + z_0\mu_2) \geq z_2^* > 0$, and $x > \frac{1}{2}$, there exists $\bar{\eta} > 0$ such that for all $\eta < \bar{\eta}$,we have $\rho_1 > \rho_2$. Hence for all $\eta < \bar{\eta}$, $\mu_2 = 0$. ■

Lemma 5 Suppose $C = \{\alpha_1, \alpha_2, \alpha_3\}$ with $\frac{1}{2} < \alpha_1$, $\alpha_2 = 1 - \alpha_1$, and $\alpha_3 \in (\alpha_2, \alpha_1)$, and $(z_1, z_2, z_3) = ((1-\eta) z^*, \eta)$ where $\eta > 0$ and $z^* \in \Omega^2$. Then $q_{13} > 0$.

Proof. Let $\alpha_1 = x$ and $\alpha_3 = y \in (1-x, x)$. Lemma 2 implies that $\mu_1 > 0$. Following the same reasoning as in Lemma 4, we can show that $\mu_2 = 0$ and $\lim_{\eta \to 0} \mu_3 = 0$. The payoff to rational players choosing x and y respectively are

$$\begin{aligned} \rho_1 &= (1-x)(z_1 + z_0\mu_1) + xz_2 + \rho_{13}(z_3 + z_0\mu_3) \\ \rho_3 &= \rho_{31}(z_1 + z_0\mu_1) + \left(\frac{y}{1-x+y}\right) z_2 + (z_3 + z_0\mu_3) \min\left\{1 - y, \frac{1}{2}\right\} \end{aligned}$$

If $q_{13} = 0$ then $\rho_{31} = 1 - x$ and, since $\rho_{13} \geq 1 - y$ and $y/(1-x+y) < x$, we have $\rho_1 > \rho_3$. Hence $q_{13} > 0$. ■

Proof of Theorem 1. Suppose that there exists $\alpha_i \in C$ such that $\alpha_j = 1 - \alpha_i \notin C$. Since $1 - \alpha_i \notin C$ there exists a mutant demand $\alpha_m > \alpha_i$ such that the set of demands in $C' = C \cup \{\alpha_m\}$ that are compatible with α_m is the same as the set of demands in C' that are compatible with α_i. Let $S = \{\alpha_k \in C' \mid \alpha_i + \alpha_k \leq 1\}$ denote this set. Consider the state

$z = ((1-\eta) z^*, \eta)$ where η is the population share of a mutant demanding α_m. Consider first the case $\mu_m = 0$. From Lemma 1, this implies that $\mu_i = 0$. The payoffs to the two behavioral types are

$$\begin{aligned}
\pi_m (C', z) &= \sum_{\alpha_k \notin S} z_0 \mu_k \alpha_m + \sum_{\alpha_k \in S} (z_k + z_0 \mu_k) \frac{\alpha_m}{\alpha_m + \alpha_k}, \\
\pi_i (C', z) &= \sum_{\alpha_k \notin S} z_0 \mu_k \alpha_i + \sum_{\alpha_k \in S} (z_k + z_0 \mu_k) \frac{\alpha_i}{\alpha_i + \alpha_k}.
\end{aligned}$$

Hence $\pi_m > \pi_i$ and the mutant can invade.

Consider next the case $\mu_m > 0$. The payoffs rational players demanding α_m and α_i are given, respectively, by

$$\begin{aligned}
\rho_m &= \sum_{\alpha_k \notin S} z_0 \mu_k q_{km} \alpha_m + \sum_{\alpha_k \notin S} (z_k + z_0 \mu_k (1 - q_{km})) (1 - \alpha_k) + \sum_{\alpha_k \in S} (z_k + z_0 \mu_k) \frac{\alpha_m}{\alpha_m + \alpha_k} \\
\rho_i &= \sum_{\alpha_k \notin S} z_0 \mu_k q_{ki} \alpha_i + \sum_{\alpha_k \notin S} (z_k + z_0 \mu_k (1 - q_{ki})) (1 - \alpha_k) + \sum_{\alpha_k \in S} (z_k + z_0 \mu_k) \frac{\alpha_i}{\alpha_i + \alpha_k}
\end{aligned}$$

Since $\mu_m > 0$ we must have $\rho_m \geq \rho_i$. This, together with the above, implies that there exists k such that $q_{km} \leq q_{ki}$. From Lemma 3, we therefore have $q_{km} \leq q_{ki}$ for all $k \in C$. Define $\tilde{p}_{ki}$ as follows:

$$\tilde{p}_{ki} = \frac{z_k}{z_k + z_0 \mu_k (1 - q_{ki})}.$$

This is the probability that a player demanding α_k is behavioral conditional on her not having conceded to demand α_i at time 0. Since $q_{km} \leq q_{ki}$ for all k, we have $\tilde{p}_{ki} \leq \tilde{p}_{km}$.

Let $\tilde{T}_{ki} = -(\log \tilde{p}_{ki}) / \lambda_{ki}$ denote the time at which such a player is first known to be behavioral. The expected payoff to a rational player demanding α_i when confronting a player demanding α_k who does not immediately concede is $1 - \alpha_k$. This is the same payoff that would be obtained by a rational player conceding with probability 1 at time $\tilde{T}_{ki}$ (since concession at this time is in the support of the equilibrium strategy). Hence

$$1 - \alpha_k = \left(\frac{z_0 \mu_k}{z_k + z_0 \mu_k} \right) \theta_{ik} + \left(\frac{z_k}{z_k + z_0 \mu_k} \right) (1 - \alpha_k) e^{-r \tilde{T}_{ki}}.$$

Here θ_{ik} is the expected payoff of a behavioral type i who encounters a rational player demanding k who does not immediately concede. Similarly

$$1 - \alpha_k = \left(\frac{z_0 \mu_k}{z_k + z_0 \mu_k} \right) \theta_{mk} + \left(\frac{z_k}{z_k + z_0 \mu_k} \right) (1 - \alpha_k) e^{-r \tilde{T}_{km}}.$$

From (1) and (2),

$$e^{-r\tilde{T}_{ki}} = \tilde{p}_{ki}^{\frac{\alpha_k+\alpha_i-1}{1-\alpha_i}} < \tilde{p}_{km}^{\frac{\alpha_k+\alpha_i-1}{1-\alpha_m}} = e^{-r\tilde{T}_{km}},$$

where the inequality follows from the fact that $\alpha_m > \alpha_i$ and $\tilde{p}_{ki} \leq \tilde{p}_{km}$. Hence $\theta_{mk} > \theta_{ik}$ and the mutant earns a greater expected payoff against all opponents with demands outside S. Since the mutant also does strictly better against demands in S, the state z^* is invadable. ■

Proof of Theorem 2

That the result holds when $x \notin (x_l, x_h)$ has been proved in the text. Accordingly, suppose that $x \in (x_l, x_h)$ and $C = \{\alpha_1, \alpha_2\}$ where $\alpha_1 = x \in (\frac{1}{2}, 1)$ and $\alpha_2 = 1 - x$. As in Lemma 4, $\mu_2 = 0$ (rational players will never imitate the less aggressive behavioral type) and the payoffs to the three types are

$$\begin{aligned}
\pi_0 &= (1-x)(1-z_2) + xz_2 - \varepsilon \\
\pi_1 &= (1-x)(1-z_2) + xz_2 - z_1(1-x)e^{-rT} \\
\pi_2 &= (1-x)(1-z_2) + \frac{1}{2}z_2
\end{aligned}$$

where

$$e^{-rT} = \left(\frac{z_1}{1-z_2}\right)^{\left(\frac{2x-1}{1-x}\right)} \tag{9}$$

All payoffs must be equal at any state z^* satisfying (3). Equality of π_0 and π_1 implies

$$z_1^*(1-x)e^{-rT^*} = \varepsilon, \tag{10}$$

where

$$e^{-rT^*} = \left(\frac{z_1^*}{1-z_2^*}\right)^{\left(\frac{2x-1}{1-x}\right)}. \tag{11}$$

Equality of π_0 and π_2 implies

$$z_2^* = \frac{2\varepsilon}{2x-1}. \tag{12}$$

We first show that this state satisfies the stability condition (3), and then that it is uninvadable.

Stability

Let $s^* = (C, z^*)$ and $s = (C, z)$ for some $z \in \Omega^2$. Since $\pi_0(s^*) = \pi_1(s^*) = \pi_2(s^*)$, we have $f(s^*, s^*) = f(s, s^*)$ for all $z \in \Omega^2$. Hence we need to show that $f(s^*, s) > f(s, s)$ for all $s \neq s^*$. Let $\Delta(z) = f(s^*, s) - f(s, s)$. Then

$$\begin{aligned}
\Delta(z) &= \sum_{i=0}^{n} (z_i^* - z_i)\,\pi_i(s) \\
&= (z_0^* - z_0)\,\pi_0(s) + (z_1^* - z_1)\,\pi_1(s) + (z_2^* - z_2)\,\pi_2(s) \\
&= (z_1 + z_2 - z_1^* - z_2^*)\,\pi_0(s) + (z_1^* - z_1)\,\pi_1(s) + (z_2^* - z_2)\,\pi_2(s)
\end{aligned}$$

Hence

$$\begin{aligned}
\Delta(z) = (z_1 + z_2 - z_1^* - z_2^*)\,&((1-x)(1-z_2) + xz_2 - \varepsilon) \\
&+ (z_1^* - z_1)\left((1-x)(1-z_2) + xz_2 - z_1(1-x)\,e^{-rT}\right) \\
&\qquad + (z_2^* - z_2)\left((1-x)(1-z_2) + \frac{1}{2}z_2\right)
\end{aligned}$$

Simplifying yields $\Delta(z) = \Delta_1(z) + \Delta_2(z)$ where

$$\begin{aligned}
\Delta_1(z) &= (1-x)(z_1^* - z_1)\left(z_1^* e^{-rT^*} - z_1 e^{-rT}\right), \\
\Delta_2(z) &= \frac{1}{2}(2x-1)(z_2^* - z_2)^2.
\end{aligned}$$

Clearly $\Delta_2(z) > 0$ for all z such that $z_2 \neq z_2^*$. To prove the result, we need to show that $\Delta_1(z) > 0$ for all z such that $z_1 \neq z_1^*$.

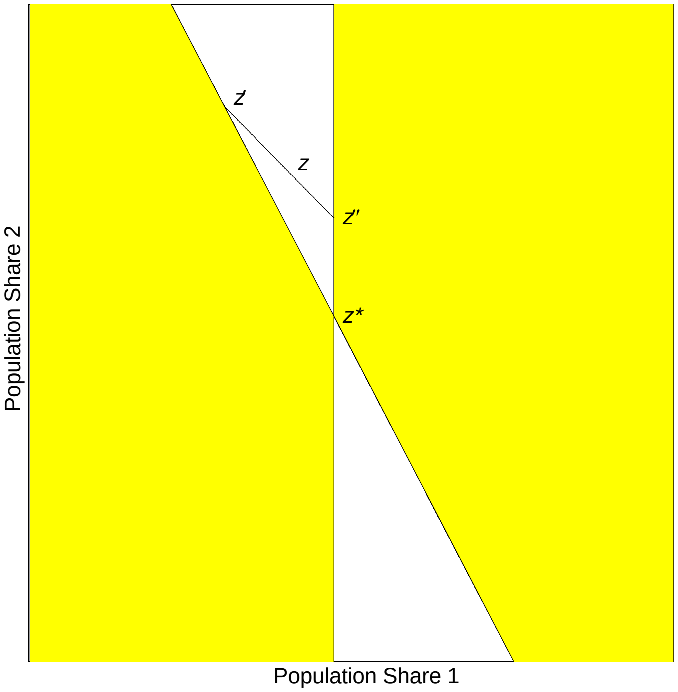

Figure 3. Steps in the Proof of Theorem 2

Consider the set of points K defined as follows

$$K = \left\{ z \in \Omega^2 \left| \frac{z_1}{1 - z_2} = k^* \right. \right\}$$

where $k^* = z_1^* / (1 - z_2^*) = z_1^* / (z_1^* + z_0^*) < 1$. This defines a straight line in z_1–z_2 space. The line passes through z^* and has slope $1/k^* > 1$ (see Figure 3). From (9) and (11), $e^{-rT} = e^{-rT^*}$ at all such points. Hence for all $z \in K$ with $z_1 \neq z_1^*$, we have

$$\Delta_1(z) = (1 - x)(z_1^* - z_1)^2 e^{-rT^*} > 0.$$

Now consider any point z such that $z_1 > z_1^*$ and $z_2 > \tilde{z}_2$ where $(z_1, \tilde{z}_2) \in K$. At any such point $z_1/(1 - z_2) > z_1/(1 - \tilde{z}_{22}) = z_1^*/(1 - z_2^*)$. Hence $e^{-rT} > e^{-rT^*}$ and $\Delta_1(z) > 0$. An analogous argument shows that for any point (z_1, z_2) such that $z_1 < z_1^*$ and $z_2 < \tilde{z}_2$ where $(z_1, \tilde{z}_2) \in K$, we also have $\Delta_1(z) > 0$. Hence $\Delta_1(z) > 0$ for all $z_1 \neq z_1^*$ when z lies within the shaded regions depicted in Figure 3.

Now consider any z such that $z_1 < z_1^*$ and $z_2 > \tilde{z}_2$ where $(z_1, \tilde{z}_2) \in K$, such as point z in Figure 3. For any such point there exist points z' and z'' such that $z_0 = z_0' = z_0''$, $(z_1', z_2) \in K$, and $z_1'' = z_1^*$ as shown in Figure 3. From (9) and the fact that $1 - z_2 = z_1 + z_0$, we have

$$\Delta_1(z) = (1 - x)(z_1^* - z_1) \left(z_1^* e^{-rT^*} - z_1 \left(\frac{z_1}{z_1 + z_0} \right)^{\left(\frac{2x-1}{1-x} \right)} \right).$$

Since z_0 is constant at all points on the line connecting z' to z'', $\Delta_1(z)$ decreases monotonically as one moves along this line from z' to z''. Since $z' \in K$, we have $\Delta_1(z') > 0$ and since $z_1'' = z_1^*$, we have $\Delta_1(z'') = 0$. Hence $\Delta_1(z) > 0$ for all z satisfying $z_1 < z_1^*$ and $z_2 > \tilde{z}_2$ where $(z_1, \tilde{z}_2, 1 - z_1 - \tilde{z}_2) \in K$. An analogous argument establishes that $\Delta_1(z) > 0$ for all z satisfying $z_1 > z_1^*$ and $z_2 < \tilde{z}_2$ where $(z_1, \tilde{z}_2) \in K$. Hence $\Delta(z) > 0$ for all $z \neq z^*$, proving that z^* satisfies (3).

Uninvadability

Consider a population composition $z = ((1 - \eta) z^*, \eta)$ where the mutant population share is η and the mutant demand is $y \notin C$. We consider three cases in turn.

Case 1: Suppose $y \in (0, 1 - x)$. In this case both y and $1 - x$ are compatible with all demands and the more aggressive demand $1 - x$ will yield a strictly greater payoff in all interactions for all η. Moreover, the difference $\pi_2 - \pi_3$ is bounded away from zero. Since $\lim_{\eta \to 0} \pi_0 = \lim_{\eta \to 0} \pi_1 = \lim_{\eta \to 0} \pi_2$, there exists $\bar{\eta}$ such that $\pi_3 < \min\{\pi_0, \pi_1, \pi_2\}$ for all $\eta < \bar{\eta}$.

Case 2: Suppose that $y \in (x, 1)$. This is the case analyzed in Lemma 4. As shown there, both aggressive types will be imitated by rational players, the passive type will not be imitated when η is sufficiently small, and there are no immediate concessions to aggressive demands. The expected payoff to a rational player choosing x and encountering demand x is $1 - x$. This must also be the payoff to a player who concedes with probability 1 at time $T_{11} = (-\log p_1)/\lambda_{11}$, since this action is in the support of the rational player's strategy. Hence

$$1 - x = \left(\frac{z_0\mu_1}{z_1 + z_0\mu_1}\right)\theta_{11} + \left(\frac{z_1}{z_1 + z_0\mu_1}\right)(1 - x)\, e^{-rT_{11}}$$

where θ_{11} is the payoff to a behavioral type who demands x and confronts a rational type also demanding x. Against an opponent demanding x, such a behavioral type therefore has expected payoff

$$\pi_{11} = \left(\frac{z_0\mu_1}{z_1 + z_0\mu_1}\right)\theta_{11} = (1 - x) - \left(\frac{z_1}{z_1 + z_0\mu_1}\right)(1 - x)\, e^{-rT_{11}}$$

Similarly, against an opponent demanding x, a behavioral type demanding y has expected payoff

$$\pi_{31} = (1 - x) - \left(\frac{z_1}{z_1 + z_0\mu_1}\right)(1 - x)\, e^{-rT_{13}}$$

Hence

$$\begin{aligned} \pi_1 &= (z_1 + z_0\mu_1)(1 - x) - z_1(1 - x)\, e^{-rT_{11}} + \eta\pi_{12} + z_2 x \\ \pi_3 &= (z_1 + z_0\mu_1)(1 - x) - z_1(1 - x)\, e^{-rT_{13}} + \eta\pi_{32} \end{aligned}$$

and

$$\begin{aligned} \lim_{\eta\to 0}(\pi_1 - \pi_3) &= z_2^* x - z_1^*(1 - x)\, e^{-rT^*} + z_1^*(1 - x)\lim_{\eta\to 0} e^{-rT_{13}} \\ &= \varepsilon\left(\frac{x}{x - \frac{1}{2}} - 1\right) + z_1^*(1 - x)\lim_{\eta\to 0} e^{-rT_{13}} \\ &= \frac{\varepsilon}{2x - 1} + z_1^*(1 - x)\lim_{\eta\to 0} e^{-rT_{13}} \geq \frac{\varepsilon}{2x - 1} \end{aligned}$$

which is bounded away from zero. Since $\lim_{\eta\to 0}\pi_0 = \lim_{\eta\to 0}\pi_1 = \lim_{\eta\to 0}\pi_2$, there exists $\bar{\eta}$ such that $\pi_3 < \min\{\pi_0, \pi_1, \pi_2\}$ for all $\eta < \bar{\eta}$.

Case 3: Suppose $y \in (1 - x, x)$. This is the case examined in Lemma 5, from which we know that $q_{13} > 0$, $\mu_1 > 0 = \mu_2$ and $\lim_{\eta\to 0}\mu_3 = 0$. Using (12),

$$\begin{aligned} \lim_{\eta\to 0}\rho_1 &= (1 - x)(1 - z_2^*) + x z_2^* = 1 - x + 2\varepsilon \\ \lim_{\eta\to 0}\rho_3 &= (1 - z_2^*)(y q_{13}^* + (1 - x)(1 - q_{13}^*)) + \left(\frac{y}{1 - x + y}\right) z_2^* \end{aligned}$$

where $q_{13}^{*} = \lim_{\eta \to 0} q_{13}$.

Suppose that there exists $\tilde{\eta}$ such that for all $\eta < \tilde{\eta}$ we have $\mu_3 = 0$ (the remaining case is treated below). then $q_{13}^{*} = z_0^{*} / \left(1 - z_2^{*}\right)$ and the payoff to the behavioral mutant satisfies

$$\lim_{\eta \to 0} \pi_3 = \lim_{\eta \to 0} \rho_3 - z_1^{*} \left(1 - x\right).$$

Since $\mu_3 = 0$ we must have $\lim_{\eta \to 0} \rho_3 \leq \lim_{\eta \to 0} \rho_1 = 1 - x + 2\varepsilon$. Hence the above implies

$$\lim_{\eta \to 0} \pi_3 \leq 1 - x + 2\varepsilon - z_1^{*} \left(1 - x\right)$$

From (10) we have $z_1^{*} \left(1 - x\right) = \varepsilon e^{rT} > \varepsilon$ so $\lim_{\eta \to 0} \pi_3 < 1 - x + \varepsilon = \lim_{\eta \to 0} \pi_1$. Moreover, since T is finite $\lim_{\eta \to 0} \left(\pi_1 - \pi_3\right)$ is bounded away from zero. Since $\lim_{\eta \to 0} \pi_0 = \lim_{\eta \to 0} \pi_1 = \lim_{\eta \to 0} \pi_2$, there exists $\bar{\eta}$ such that $\pi_3 < \min\{\pi_0, \pi_1, \pi_2\}$ for all $\eta < \bar{\eta}$.

Finally consider the case in which $\mu_3 > 0$ for all $\eta > 0$. Let $\tilde{p}_1$ be the probability that a player is behavioral conditional on the fact that she has chosen x and has not conceded immediately to an opponent choosing y. Conditional on her opponent choosing x and not conceding immediately, the expected payoff to a rational player choosing y is $1 - x$. This is the same as the payoff to a player who concedes with probability 1 at the time $\tilde{T}$ when the opponent is revealed to be behavioral, where $\tilde{T} = (-\log \tilde{p}_1)/\lambda_{13}$. Let $\tilde{p}_1^{*}$ and $\tilde{T}^{*}$ denote $\lim_{\eta \to 0} \tilde{p}_1$ and $\lim_{\eta \to 0} \tilde{T}$ respectively. Hence

$$1 - x = \left(1 - \tilde{p}_1^{*}\right) \theta_{31} + \tilde{p}_1^{*} \left(1 - x\right) e^{-r\tilde{T}^{*}}$$

where θ_{31} is the payoff to a player who concedes with probability 1 at the time $\tilde{T}^{*}$ conditional on a match with a rational player. A behavioral player choosing y, conditional on her opponent choosing x and not conceding immediately, therefore gets $\left(1 - \tilde{p}_1^{*}\right) \theta_{31} = 1 - x - \tilde{p}_1^{*} \left(1 - x\right) e^{-r\tilde{T}^{*}}$. In all other circumstances the payoffs to behavioral and rational players choosing y are identical. The payoff to behavioral players therefore satisfies

$$\lim_{\eta \to 0} \pi_3 = \lim_{\eta \to 0} \rho_3 - \left(1 - z_2^{*}\right) \left(1 - q_{13}^{*}\right) \left(1 - x - \tilde{p}_1^{*} \left(1 - x\right) e^{-r\tilde{T}^{*}}\right)$$

Since $\mu_3 > 0$, we have $\lim_{\eta \to 0} \rho_3 = \lim_{\eta \to 0} \rho_1 = 1 - x + 2\varepsilon$. Hence

$$\lim_{\eta \to 0} \pi_3 = 1 - x + 2\varepsilon - \left(1 - z_2^{*}\right) \left(1 - q_{13}^{*}\right) \tilde{p}_1^{*} \left(1 - x\right) e^{-r\tilde{T}^{*}} \tag{13}$$

We claim that $\left(1 - z_2^{*}\right) \left(1 - q_{13}^{*}\right) \tilde{p}_1^{*} \left(1 - x\right) e^{-r\tilde{T}^{*}} > \varepsilon$. To see this, recall from (10) that $z_1^{*}(1 - x)e^{-rT} = \varepsilon$. Hence it is sufficient to show that (i) $e^{-r\tilde{T}^{*}} > e^{-rT}$ and (ii) $\left(1 - z_2^{*}\right) \left(1 - q_{13}^{*}\right) \tilde{p}_1^{*} \geq z_1^{*}$. First we prove that (i) holds. Recall that $\tilde{T}^{*} = (-\log \tilde{p}_1^{*})/\lambda_{13}$ and $T = (-\log p_1)/\lambda_{11}$.

It is trivially the case that $\tilde{p}_1^* > p_1$ (since behavioral types never concede) and $y < x$ implies $\lambda_{13} > \lambda_{11}$ from (1). Hence $\tilde{T}^* < T$ which proves (i). To prove (ii), observe that $\tilde{p}_1^* = z_1^* / (1 - z_2^* - q_{13})$. Hence

$$(1 - z_2^*)(1 - q_{13}^*)\,\tilde{p}_1^* = \left(\frac{(1 - z_2^*)(1 - q_{13}^*)}{1 - z_2^* - q_{13}} \right) z_1^*$$

The expression in parenthesis is increasing for all $q_{13}^* \in (0, z_0^*)$ with a minimum value of 1. Hence (ii) holds, which proves the claim that $(1 - z_2^*)(1 - q_{13}^*)\,\tilde{p}_1^*(1 - x)\,e^{-r\tilde{T}^*} > \varepsilon$. Using (13) together with this claim yields $\lim_{\eta \to 0} \pi_3 < 1 - x + \varepsilon = \lim_{\eta \to 0} \pi_i$ for all $i = 0, 1, 2$. Hence there exists $\bar{\eta}$ such that $\pi_3 < \min\{\pi_0, \pi_1, \pi_2\}$ for all $\eta < \bar{\eta}$. ■

References

[1] Abreu, D. and F. Gul (2000). “Bargaining and Reputation.” *Econometrica* 68: 85–117.

[2] Banerjee, A.V. and J.W. Weibull (1995). “Evolutionary Selection and Rational Behavior.” In A Kirman and M. Salmon, eds., Learning and Rationality in Economics. Oxford: Blackwell.

[3] Conlisk, J. (1980). “Costly Optimizers Versus Cheap Imitators.” *Journal of Economic Behavior and Organization* 1: 275–93.

[4] Dekel, E. and S. Scotchmer (1992). “On the Evolution of Optimizing Behavior.” *Journal of Economic Theory* 57: 392–406.

[5] Fudenberg, D. and E. Maskin (1986). The Folk Theorem in Repeated Games with Discounting and with Incomplete Information. *Econometrica* 54, 533–554.

[6] Hofbauer, J. and K. Sigmund (1988). *The Theory of Evolution and Dynamical Systems.* Cambridge: Cambridge University Press.

[7] Guttman J.M. (1996). “Rational Actors, Tit-for-tat Types, and the Evolution of Cooperation.” *Journal of Economic Behavior and Organization* 29: 27–56.

[8] Kreps, D.M., P. Milgrom, J. Roberts and R. Wilson (1982). “Rational Cooperation in the Finitely Repeated Prisoners’ Dilemma.” *Journal of Economic Theory* 27: 245–252.

[9] Kreps, D.M. and R. Wilson (1982). “Reputation and Imperfect Information.” *Journal of Economic Theory* 27: 253–279.

[10] Maynard Smith, J. and G.R. Price (1973). “The Logic of Animal Conflict.” *Nature* 246: 15–18.

[11] McElvey, R.D. and T.R. Palfrey (1992). “An Experimental Study of the Centipede Game.” *Econometrica* 60: 803–836.

[12] Milgrom, P. and J. Roberts (1982). “Predation, Reputation, and Entry Deterrence.” *Journal of Economic Theory* 27: 280–312.

[13] Simon, H.A. (1978). “Rationality as Process and Product of Thought.” *American Economic Review* 68: 1–16.

[14] Stahl, D.O. (1993). “Evolution of Smart$_n$ Players”. *Games and Economic Behavior* 5: 604–17.