Evolutionary Game Theory and the Prisoner's Dilemma: a survey

Douglas G. Kelly
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
October 2008

Contents

0. Introduction. This is an expository survey, intended to introduce, at the upper undergraduate level, enough game theory, evolutionary game theory, and evolutionary population dynamics to begin a study of the ecology of populations of individuals playing various strategies in games involving competition and cooperation. Some familiarity with linear algebra, differential equations, and elementary probability is assumed.

Sections 1 through 5 introduce the Prisoner's Dilemma, which is a particular example of a symmetric 2×2 game, and the Iterated Prisoner's Dilemma (IPD), in which competitious involving finite sets of strategies can be viewed as symmetric $n \times n$ games.

Sections 6 through 8 introduce some of the ideas of game theory, focusing on mixed strategies viewed as populations of pure-strategy players. We discuss some equilibrium properties of mixed strategies – in particular, Nash equilibrium and evolutionary stability – that appear meaningful for studying the evolution of such populations.

In Sections 9 and 10 we introduce the *replicator dynamics*, a system of ordinary differential equations whose orbits represent population evolution. We show (in some cases without proof; the proofs are in another paper to come) all the connections among

the equilibrium properties from Sections 6 through 8 and the various types of stationary points of the replicator dynamics. These connections are summarized in Table 10.1.

Our exposition of the theory in Sections 6 through 10 is based heavily on the coverage in three important books on the subject, namely those by Hofbauer and Sigmund [1998], Osborne and Rubinstein [1994], and Weibull [1997].

In Secion 11 we cover the simplest special case, 2×2 symmetric games, classifying and describing all possible types of population dynamics in terms of the parameters of the game. In Section 12 we apply these results to some two-strategy populations for the IPD.

For 3×3 games, such a complete categorization is sufficiently complicated to be not only infeasible but also unenlightening. In Section 13 we consider a fairly narrow subclass of the 3×3 games and describe completely all types of dynamics for these. Though narrow, this subclass includes the games determined by three particular strategies for the IPD, suggested in a paper of Boyd and Lorberbaum [1987]. In Section 14 we examine the dynamics of these three strategies. We confirm Boyd and Lorberbaum's observation that the celebrated strategy *Tit for Tat* is in fact not evolutionarily stable; under some parameter settings it is driven to extinction by the combination of a more "forgiving" strategy, *Tit for Two Tats*, and a less "nice" strategy, *Suspicious Tit for Tat*.

1. The Prisoner's Dilemma. We start with a game invented in 1950 by Merrill Flood and Melvin Dresher of RAND. The history of this game, and many of its ramifications, are engagingly told in William Poundstone's book *Prisoner's Dilemma* [1993[. Albert W. Tucker named the game the ***Prisoner's Dilemma*** and told something like the following story to motivate it.

> Two people have been arrested in connection with a crime and are being held, unable to communicate with each other. The prosecuting attorney talks to each of them separately, saying: "This crime carries a maximum sentence of six years in prison. If neither of you chooses to confess to it, I will have some trouble prosecuting you, but I can still get a three-year sentence for each of you. If one of you confesses and the other doesn't, the confessor's sentence will be reduced to one year, while the other will get the full six years. If you both confess, you will each get five years."

If you were one of these prisoners, what would you do? You have a choice of two actions: *Cooperate* with your partner in crime and refuse to confess, or *Defect* and turn State's evidence. The payoff (sentence reduction, in years) to either prisoner depends on what both prisoners do, according to the following table.

		B's action: Cooperate	B's action: Defect
A's action	Cooperate	3 for A, 3 for B	0 for A, 5 for B
	Defect	5 for A, 0 for B	1 for A, 1 for B

The dilemma, once thought of as a paradox, is this. Each prisoner looks at the payoff table and notices that, regardless of what the other prisoner does, it is better to defect. If they could talk to each other, agree to cooperate, and trust each other, they could get three-year reductions. But they can't, and the result is that both defect and get only one-year reductions, simply because of their inability to communicate and lack of trust.

As applied to two criminals, this story's moral may be ambiguous. But real life is full of similar situations – "games" involving two or more players, in which defecting takes the form of "free riding." Garrett Hardin [1968] wrote about the "Tragedy of the Commons." Hardin's parable is of common grazing land, maintained and used by a collection of keepers of livestock. As long as they all "cooperate" and keep their herds at moderate sizes, all can use the land comfortably. But anyone who "defects" and unduly increases his herd size will gain. If many choose to defect, all suffer in the long run. Yet in the optimal "all cooperate" state, there is a strong incentive for any individual to defect.

Analogous situations abound in human society. For example:

> Why should public-radio listeners bother to contribute to their local stations? Your contribution is a noticeable expense for you but its absence would have little effect on the station.

Late at night in a subway station, with no one around, why not jump over the turnstile to avoid paying? It saves you a fare and costs the system almost nothing.

In traffic, where two lanes are narrowing into one, the flow will be smoother for everyone if drivers alternate entering the single lane, but a defector saves some time at cost to the others.

Why should you vote, since it's inconvenient for you and you can be virtually certain that your vote will not affect the outcome?

In each case, the benefit to you of defecting, or free riding, is greater than its negative effect on the world. So people defect; and when many defect, the world is worse than it might be.

On the other hand, many of us do cooperate, and as a result the world isn't as bad as it could be. The question is: where does cooperation come from, in humans and in some animal species? Where do we get our moral sense of wanting to be part of the solution and not part of the problem? Religious answers may be sufficient in some contexts, but, like “intelligent design,” they don't get to the mechanisms by which deities work in the world. Scientists, religious or not, want to understand these mechanisms.

The Prisoner's Dilemma is a minimal, and clearly insufficient, model for interactions involving competition and cooperation among humans. What do we need to add to the model so that cooperative behavior finds a place?

There is a twofold partial answer.

First, the relevant game is not the Prisoner's Dilemma, but the so-called ***Iterated Prisoner's Dilemma***, in which players will play series of Prisoner's Dilemma games against each other and may discover the advantages of cooperation and the possibility of punishing defectors.

Second, extending game theory to ***evolutionary game theory*** provides a way to see how populations of players, using different strategies, may attain stability with a mix of different degrees of cooperativeness.

2. The Prisoner's Dilemma as a two-player game. A general two-player game takes place between two people, a "row player" whom we will call Rhoda, and a "column player," called Colleen. The game has a payoff table of the form

		Colleen's actions			
		C_1	C_2	$\cdots$	C_n
Rhoda's actions	R_1	a_{11}, b_{11}	a_{12}, b_{12}	$\cdots$	a_{1k}, b_{1k}
	R_2	a_{21}, b_{21}	a_{21}, b_{21}	$\cdots$	a_{2k}, b_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	R_k	a_{n1}, b_{n1}	a_{n2}, b_{n2}	$\cdots$	a_{nk}, b_{nk}

where $a_{ij} \in \mathbb{R}$ is the payoff to Rhoda, and $b_{ij} \in \mathbb{R}$ is the payoff to Colleen, when they take actions R_i and C_j, respectively. The two $n \times k$ matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ determine the game.

The game is ***symmetric*** if $n = k$, $R_i = C_i$ $(i = 1\ldots, n)$, and $B = A^T$. Note that the matrix A of a symmetric game is not necessarily a symmetric matrix. If A does happen to be symmetric, the game is called *doubly symmetric*.

We will consider only symmetric two-player games in these notes.

A ***Prisoner's Dilemma*** (PD) is defined to be a symmetric two-player game with two actions, denoted C (Cooperate) D (Defect) instead of R_1 and R_2, and payoff matrices

	C	D
C	R, R	S, T
D	T, S	P, P

generalizing the original

	C	D
C	$3, 3$	$0, 5$
D	$5, 0$	$1, 1$

,

with $S < P < R < T$ and $R > \frac{1}{2}(S + T)$. These four letters are chosen so that we can remember

R as the Reward for mutual cooperation,
S as the Sucker's penalty,
T as the Temptation to defect, and
P as the Penalty for mutual defection.

The second condition, $R > \frac{1}{2}(S + T)$, will be important when we consider two people playing repeatedly against each other. Then R is the average gain per play of people who cooperate with each other, while $\frac{1}{2}(S + T)$ is the average gain per play of people who alternately cooperate and defect, out of phase with each other. We will see the importance of this below.

Customarily we will define a symmetric game by giving only the matrix A, since $B = A^T$. Thus we would represent the Prisoner's Dilemma with

$$A = \begin{bmatrix} R & S \\ T & P \end{bmatrix} \text{or, in the original example,} \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix},$$

remembering that the entries represent payoffs to Rhoda, the row player. We will speak of the parameters $R = 3$, $S = 0$, $T = 5$, $P = 1$ as the "Axelrod numbers," for reasons that we will see below.

3. The Iterated Prisoner's Dilemma. Suppose two players repeatedly play the PD against each other. They can decide on each play whether to cooperate or defect, based on the history of the game so far. Here is a possible outcome of a ten-play round between two players, X and Y.

play	1	2	3	4	5	6	7	8	9	10	total
X's action (payoff)	C (3)	C (0)	D (1)	C (3)	C (3)	D (5)	C (0)	C (0)	C (3)	C (0)	18
Ys action (payoff)	C (3)	D (5)	D (1)	C (3)	C (3)	C (0)	D (5)	D (5)	C (3)	D (5)	33

If they had both cooperated on every play, they would each have got 30 points in all. But Y was able to improve on that by being willing to defect, taking advantage of X's cooperativeness.

At this point it is helpful to think for a while about how one might play the IPD effectively, assuming that one's opponent, like oneself, is bright and analytic.

The IPD became famous beginning in the late 1970s, when Robert Axelrod, a political scientist then at the University of Michigan, conducted a series of tournaments. Axelrod asked various researchers to contribute strategies, and then ran computer tournaments in which each submitted strategy was pitted against every other submitted strategy, as well as against a copy of itself, along with some random strategies in which the actions are Bernoulli sequences. Axelrod's results are summarized in a pair of books, *The Evolution of Cooperation* [1984] and *The Complexity of Cooperation* [1997].

Here are some of the simpler strategies submitted to Axelrod's tournaments.

Tit for Tat: Cooperate on the first play; on subsequent moves do whatever opponent did on the previous play.

Always Cooperate.

Always Defect.

Grudger ("Massive Retaliation"): Cooperate until opponent defects, then defect on every play thereafter.

Suspicious Tit for Tat: Defect on the first play; on subsequent plays do whatever opponent did on the previous play.

Tit for Two Tats: Cooperate until opponent has defected twice in succession; then defect until the opponent cooperates again; start over.

Prober: Same as Tit for Tat, but wherever a C is called for, D with some small probability p (say $p = 0.1$).

Peacemaker: Same as Tit for Tat, but wherever a D is called for, C with some small probability p.

Adaptive: Begin with CCCCCCDDDDD; then on each play choose the action that has given the best average return to date. [There are obviously variants, using different initial sequences.]

Soft Grudger: Cooperate until opponent defects, then punish with DDDDCC. [Obvious variants use different sequqnces of punishment followed by forgiveness.]

Pavlov, or Win-Stay, Lose-Switch: Cooperate on the first play; after that, repeat the previous action if opponent cooperated on that play, change if opponent defected.

The most important and surprising early finding in these tournaments was the success of Tit for Tat (TFT), the invention and tournament entry of Anatol Rapoport. It is one of the simplest strategies; it looks only at the opponent's previous move and merely repeats it. It is not hard to see that TFT can never beat any opponent in head-to-head competition: at the end of a round it will either be tied with the opponent or have 5 points less. Yet it consistently ranked at or near the top against the field in Axelrod's competitions.

Competitions have been held regularly by a group whose website can be found at <http://www.prisoners-dilemma.com>. Strategies have become quite elaborate in recent years. TFT can be beaten, though not badly; it continually ranks among the most successful strategies. Another successful simple strategy is Pavlov, which strangely didn't enter the public consciousness until the early 1990s. It appears to outperform TFT, depending of course on the other strategies in the contest.

Notice that the emphasis has shifted when we think about tournaments. Instead of trying to beat opponents in single rounds of head-to-head competition, we would like a strategy that performs well against many other strategies. As we have mentioned, there are highly successful strategies like TFT that cannot beat *any* other strategy head-to-head, but are never beaten badly by any of them.

In his early work Axelrod noticed some qualities of strategies that seemed to make them successful in his tournaments. These observations are provocative, if sometimes vague, and rigorous work on them has been difficult to achieve.

One such quality is *niceness.* A ***nice*** strategy is one that is never the first to defect. Examples are TFT, Grudger, Soft Grudger, Pavlov, and several others. Suspicious Tit for Tat, Prober, Adaptive, and Always Defect are not nice.

A strategy is called ***provokable*** to the extent that it responds to defections by defecting one or more times. TFT is provokable, Tit for Two Tats less so; Always Cooperate is not provokable; Grudger is maximally provokable.

A strategy is called ***forgiving*** to the extent that it resumes cooperating at some point after an opponent's defection. TFT is quite forgiving; Always Cooperate is perfectly forgiving; Grudger is perfectly non-forgiving.

Axelrod noticed, and hypothesized, that successful strategies tend to be nice, provokable, and forgiving.

In reading Axelrod's early papers and his first book, one is struck by their qualitative viewpoint. Axelrod scarcely describes the submitted strategies, but takes care to document who submitted them, what fields of expertise and what organizations the

> submitters came from, the lengths of the submitted computer programs, and the languages the programs were written in. In later work, of course – and even in his earliest work, as noted above – Axelrod recognized the importance of trying to identify the qualities that made strategies successful.

Axelrod used the parameter values $R = 3$, $S = 0$, $T = 5$, $P = 1$ in his tournaments, and while he was not the first to use these particular values, we will refer to them as the "Axelrod numbers."

Where are we heading? At first one may approach the IPD by looking for a strategy that beats other strategies in head-to-head competition. Tournaments like Axelrod's put the emphasis not on head-to-head competition but on performing well against a field of opponents playing a wide range of strategies; it is not important to do well against individuals, or even to beat a large number of others, as it is to do better than anyone else against the field.

But when we think about the *evolution of cooperation among self-interested agents,* the emphasis shifts again. Now the goal is not to devise strategies that win in any sense, but to see what kinds of strategies can exist in stable equilibrium in a population.

Here "stable equilibrium" refers to the following scenario: in a large population of individuals, each individual plays one strategy consistently in encounters against a large number of randomly-chosen opponents. Strategies gain "fitness points" proportional to their average gains in these encounters. Individuals reproduce, passing their strategies to their progeny, and the number of progeny of an individual playing a given strategy is proportional to the number of fitness points she earned. A population is in stable equilibrium if the population shares of the pure strategies do not change over time. As we will see, this occurs when all strategies have the same expected average gain per encounter. Sections 6 through 10 below expand on this sketch of the scenario.

4. IPD competition as a two-player game. Now we observe that, given a set of strategies, we can regard an IPD round as a single play in a symmetric two-player game, in which the actions are the strategies in the given set.

To illustrate, suppose a number of players each choose one of the three strategies Tit for Tat (TFT), Suspicious Tit for Tat (STFT), and Tit for Two Tats (TF2T). How will they do in IPD rounds against each other (including against players using the same strategy as themselves)?

Here, for example, is what happens in a round between two players, one playing TFT and the other STFT.

play	1	2	3	4	5	6	7	8	9	10	average
TFT	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	$\frac{1}{2}(S+T)$
STFT	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	$\frac{1}{2}(S+T)$

The averages will change slightly if the round has an odd number of plays, but asymptotically for long rounds each player will earn $\frac{1}{2}(S+T)$ points per round. With the Axelrod numbers this is 2.5.

Note that if TFT plays TFT, they get R per play for an average of R, and we have assumed $R > \frac{1}{2}(S+T)$. This requirement makes mutual cooperation preferable to the kind of cycle of retribution shown above.

As another example, consider a long round of TF2T against STFT.

play	1	2	3	4	$\cdots$	average
TF2T	C (S)	C (R)	C (R)	C (R)	$\cdots$	$\sim R$
STFT	D (T)	C (R)	C (R)	C (R)	$\cdots$	$\sim R$

TF2T does better against STFT than TFT does. (However, TF2T is easily exploited by other non-nice strategies, which are not present in this example.)

It is not difficult to check that the following table gives the asymptotic average scores per round for each of the strategies against itself and the others.

	TFT	TF2T	STFT
TFT	$R\,(=3)$	$R\,(=3)$	$\frac{1}{2}(S+T)\,(=2.5)$
TF2T	$R\,(=3)$	$R\,(=3)$	$R\,(=3)$
STFT	$\frac{1}{2}(S+T)\,(=2.5)$	$R\,(=3)$	$P\,(=1)$

(4.1)

It is clear that strategies for the IPD can be thought of as the actions in a symmetric two-player game. A single play of this game is an IPD round.

In general, any set of n IPD strategies determines a symmetric two-player game with n actions available to each player. The payoffs are the average winnings of the strategies against each other (and copies of themselves) in long rounds. (If the strategies have randomness in them, the payoffs will be expected average winnings.)

5. The length of the round; discounting the future. People who play the IPD soon recognize that it always pays to defect on the last play of the round. You cannot do better by cooperating, and your opponent cannot punish you by defecting subsequently. Indeed, if you believe your opponent has this same insight, you may decide that you ought to defect on the next-to-last play also, and so on. The well-known "unexpected hanging paradox" comes to mind. Whether you reason all the way to the paradox or not, any strategy that does not always defect on the last play of the round can be improved by modifying it to do so.

To avoid this, Axelrod decided to stop the IPD rounds in his tournaments at random times unknown to the players. Players engage in shorter or longer rounds, but the number of plays over many rounds is approximately constant, so that average gain per play (or expected average gain per play in the case of randomized strategies) is equivalent to total gain as a measure of success.

One way to achieve random stopping is to decide after each play to continue with probability w or to stop with probability $1-w$. If the decisions are made independently, then the number of plays equals k with probability $(1-w)w^{k-1}$ $(k = 1, 2, \dots)$.

We can then compute the expected payoffs of the various strategies against each other. In general, if g_k denotes the payoff to strategy S_1 on the k^{th} play against a given opponent strategy S_2, then the expected total payoff to S_1 in a randomly-stopped game is

$$\sum_{k=1}^{\infty}(1-w)w^{k-1}\sum_{j=1}^{k}g_j = (1-w)\sum_{j=1}^{\infty}g_j\sum_{k=j}^{\infty}w^{k-1} = \sum_{j=1}^{\infty}g_j w^{j-1}.$$

We see that this is the same as the total gain in an infinitely long IPD round, if the gain on the j^{th} play is discounted by w^{j-1}. We will refer to w as the *discount factor*.

For TFT *vs.* STFT, for example, we have as above

play	1	2	3	4	5	6	7	8	$\cdots$
TFT	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	$\cdots$
STFT	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	D (T)	C (S)	$\cdots$

and so the expected total payoffs are

$$S + Tw + Sw^2 + Tw^3 + \cdots = \frac{S+Tw}{1-w^2} \quad \text{for TFT}$$
$$\text{and } T + Sw + Tw^2 + Sw^3 + \cdots = \frac{T+Sw}{1-w^2} \text{ for STFT.}$$

The advantage enjoyed by STFT, as measured by the ratio of their payoffs, is $\frac{T+Sw}{S+Tw}$. As w increases towards 1 (i.e. the expected length of the game increases, or, equivalently, the future counts for more), the advantage enjoyed by STFT decreases. Notice that the *difference* between their total gains increases without limit. However, it is the ratio that matters when we consider population dynamics, as we will see.

One can check that with the discount factor w, the table (4.1) is replaced by

	TFT	TF2T	STFT
TFT	$\frac{R}{1-w}$	$\frac{R}{1-w}$	$\frac{S+Tw}{1-w^2}$
TF2T	$\frac{R}{1-w}$	$\frac{R}{1-w}$	$S+\frac{Rw}{1-w}$
STFT	$\frac{T+Sw}{1-w^2}$	$T+\frac{Rw}{1-w}$	$\frac{P}{1-w}$

(5.1)

It is interesting to note the effect of w on the game. Small values of w correspond to rounds that are likely to be short – or, equivalently, infinite rounds in which the future is discounted heavily. Small values of w would seem to give STFT, with its initial defection, a greater advantage over the other strategies than larger values. And this is true. Following are the values in the table for the Axelrod numbers $R = 3$, $S = 0$, $T = 5$, $P = 1$ and two values of w. We have multiplied all entries in the first matrix by 3 for easier comparison with the second. (As we will see, this has no effect on any of our considerations.)

$w = 0.25$

	TFT	TF2T	STFT
TFT	12	12	4
TF2T	12	12	3
STFT	16	18	4

$w = 0.75$

	TFT	TF2T	STFT
TFT	12	12	$8\frac{4}{7}$
TF2T	12	12	9
STFT	$11\frac{3}{7}$	14	4

The effect of w on STFT's performance is apparent.

6. Game theory: Nash equilibrium. In general, a two-player game is determined by a pair of $n \times k$ matrices, $A = [a_{ij}]$ and $B = [b_{ij}]$. As noted above, the players are usually called the "row player" and the "column player;" we call them Rhoda and Colleen for short. Rhoda has actions $R_1, \ldots, R_n$ to choose from, and Colleen has actions $C_1, \ldots, C_k$. If they choose actions R_i and C_j, then Rhoda's payoff is a_{ij} and Colleen's is b_{ji}. It is customary to display both matrices in a single table, as we did earlier:

		Colleen's actions			
		C_1	C_2	$\cdots$	C_n
Rhoda's actions	R_1	a_{11}, b_{11}	a_{12}, b_{12}	$\cdots$	a_{1k}, b_{1k}
	R_2	a_{21}, b_{21}	a_{21}, b_{21}	$\cdots$	a_{2k}, b_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	R_k	a_{n1}, v_{n1}	a_{n2}, b_{n2}	$\cdots$	a_{nk}, b_{nk}

(6.1)

In a symmetric two-player game, $n = k$, each player has the same set of actions, and $A = B^T$. Even for symmetric games, it is sometimes helpful to list both A and B. We would describe the PD, for example, by

	C	D
C	3,3	0,5
D	5,0	1,1

or generally by

	C	D
C	R,R	S,T
D	T,S	P,P

and the IPD with three strategies TFT, TF2T, and STFT and $w = 0.75$, as seen at the end of Section 5, by

	TFT	TF2T	STFT
TFT	$12, 12$	$12, 12$	$8\frac{4}{7}, 11\frac{3}{7}$
TF2T	$12, 12$	$12, 12$	$9, 14$
STFT	$11\frac{3}{7}, 8\frac{4}{7}$	$14, 9$	$4, 4$

Of course, any set of n IPD strategies will produce a symmetric two-player game with n actions. The discount factor w must be specified, and if the strategies are random, then expected payoffs are used.

Game theory originated with the attempt to answer questions like "What should players do in such a game?" or "What *will* players do in such a game?" In the Prisoner's Dilemma, played once with no collusion and no trust, they probably will, alas, each defect, and perhaps this is what they should do. In the IPD game above, it is less clear what they will do.

Here is a nonsymmetric game in which it is instructive to think about what they will or should do.

	C_1	C_2	C_3	C_4
R_1	$5,2$	$2,6$	$1,4$	$0,4$
R_2	$0,0$	$3,2$	$2,1$	$1,1$
R_3	$7,0$	$2,2$	$1,5$	$5,1$
R_4	$9,5$	$1,3$	$0,2$	$4,8$

Imagine that Rhoda, the row player, and Colleen, the column player, are going to play this game just once. They have no opportunity to discuss the matter with each other before playing. Rhoda might reason as follows: "I'd like to play R_4 in the hopes of getting that 9-point payoff. But if Colleen knows I plan to play R_4, she'll surely play C_4 for an 8-point payoff. But if she's going to play C_4, I should play R_3 (5-point payoff). And if I play R_3, her best play is C_3. But knowing that, I'd better play R_2, and if she knows that, then she'll play C_2. But then my best play is still R_2. So if we play R_2 and C_2, neither of us will have any incentive to change."

This kind of argument is a simple form of what is called ***fictitious play***. But will two players actually follow such reasoning? Notice that neither of them will be very happy with the resulting pair of actions. But what is clear is that, if the pair (R_2, C_2) were imposed as a default or a starting point, then neither player could improve her payoff by switching to another action, unless the other player switched as well. This is because C_2 is a *best reply* to R_2 and vice versa.

Generally, in a nonsymmetric two-player game as in (6.1) above, R_i is a ***best reply*** by the row player to the column player's action C_j if $a_{ij} \geq a_{rj}$ for $r = 1, 2, \ldots, n$. Similarly, C_j is a ***best reply*** by the column player to the row player's R_i if $b_{ij} \geq b_{is}$ for $s = 1, 2, \ldots, k$. And a ***Nash equilibrium*** (NE) is a pair of actions (R_i, S_j) that are best replies to each other.

R_i is a ***strict best reply*** to C_j if $a_{ij} > a_{rj}$ for $r \neq i$ and similarly for C_j to R_i. A ***strict NE*** is a NE in which the actions are strict best replies to each other.

It is easy to find all the NE (there may be more than one) for any game, using the tabular presentation of A and B. Here is the table for the game above, with the payoffs for all the best replies in boldface.

	C_1	C_2	C_3	C_4
R_1	$5,2$	$2,\mathbf{6}$	$1,4$	$0,4$
R_2	$0,0$	$\mathbf{3,2}$	$\mathbf{2},1$	$1,1$
R_3	$7,0$	$2,2$	$1,\mathbf{5}$	$\mathbf{5},1$
R_4	$\mathbf{9},5$	$1,3$	$0,2$	$4,\mathbf{8}$

Look, for example, at the entry $\mathbf{5}, 1$ corresponding to R_3 and C_4. The $\mathbf{5}$ is in boldface, because R_3 is the row player's best reply to the column player's C_4. But C_4 is not the column player's best reply to R_3.

Wee see immediately that there is one NE: the pair (R_2, C_2) of actions in which both payoffs are in bold. This is the same pair of actions that we reached by the fictitious-play argument above. But we must note that such an argument does not always lead to a NE. Some games have more than one NE, and in some there is no pair of actions that are best replies to each other.

> Another argument that may lead to NE is *iterated elimination of dominated actions*. In the game above, for example, Rhoda may notice that she has no reason ever to play R_1; it is ***dominated by*** (i.e., never better and sometimes worse than) R_3. So she (and Colleen, since she is smart too) can eliminate R_1. One can check that in this reduced game, Colleen can eliminate C_1 since C_4 dominates it. In this further reduced game, Rhoda can eliminate R_4, and then Colleen can eliminate C_4, then Rhoda can eliminate R_3, and finally Colleen can eliminate C_3. They are left with R_2 and C_2, the NE.
>
> It can be shown that iterated elimination of dominated actions never eliminates a NE; all NE will remain. But other pairs that are not NE may also remain. Indeed, sometimes there are no dominated actions in a game, so there is no elimination.

We reiterate that a NE is not necessarily optimal in any sense. In the above example, neither player would be happy with the NE (R_2, C_2), and in fact both would prefer (R_4, C_1) or (R_4, C_4). But neither player can gain by changing actions from a NE unless her opponent changes as well.

In the PD, as one may have sadly suspected, $(\text{Defect}, \text{Defect})$ is the only NE, and it is strict.

	C	D
C	$3, 3$	$0, \mathbf{5}$
D	$\mathbf{5}, 0$	$\mathbf{1}, \mathbf{1}$

or, for any PD,

	C	D
C	R, R	$S, \boldsymbol{T}$
D	$\boldsymbol{T}, S$	$\boldsymbol{P}, \boldsymbol{P}$

.

We also refer to the action D as a ***symmetric NE***: an action that is a best reply to itself. Notice that the idea of a symmetric NE makes little or no sense in nonsymmetric games, but that there can be nonsymmetric NE in symmetric games.

In the three-strategy IPD considered above, with $w = 0.75$, there are three NE, of which two are strict and the third is symmetric.

	TFT	TF2T	STFT
TFT	$\mathbf{12}, \mathbf{12}$	$12, \mathbf{12}$	$8\frac{4}{7}, 11\frac{3}{7}$
TF2T	$\mathbf{12}, 12$	$12, 12$	$\mathbf{9}, \mathbf{14}$
STFT	$11\frac{3}{7}, 8\frac{4}{7}$	$\mathbf{14}, \mathbf{9}$	$4, 4$

(6.2)

Given that only these three strategies are available, if both players play TFT, or if one

plays TF2T and the other plays STFT, neither will have any incentive to switch strategies unless the other does also.

A game need not have any NE:

	R_1	R_2
R_1	$\mathbf{1}$,-1	$-1,\mathbf{1}$
R_2	$-1,\mathbf{1}$	$\mathbf{1},-1$

This game is called "Matching Pennies." The row player wins if the actions match and the column player wins otherwise. It is a zero-sum game, in that the row player's gain is the column player's loss and vice versa.

> This example does not contradict Nash's celebrated theorem that every finite symmetric game has a NE. That theorem guarantees a *mixed-strategy* NE, which we will cover in the next section. In this section we are talking only about *pure-strategy* NE.

Even a symmetric game need not have any NE. The following game is "Rock, Paper, Scissors," in which Rock defeats (breaks) Scissors, Scissors defeats (cuts) paper, and Paper defeats (covers) Rock.

	Rock	Paper	Scissors
Rock	$0,0$	$-1,\mathbf{1}$	$\mathbf{1}$,-1
Paper	$\mathbf{1}$,-1	$0,0$	$-1,\mathbf{1}$
Scissors	$-1,\mathbf{1}$	$\mathbf{1}$,-1	$0,0$

One can also see easily that in a symmetric game, if (R_i, R_j) is a NE, then so is (R_j, R_i).

The considerations above are not dynamic in any sense. They have to do with one play (it is irrelevant here that one IPD play comes from many plays of a different game), with no prior history or contact between the players; and they give no indication of how any combination of plays may be arrived at.

To put dynamic considerations into models of competition and cooperation, we need to begin with the notion of a mixed strategy. This will be defined next as a probability distribution over the set of actions available to a player. In this context the single actions are called *pure strategies*.

7. Mixed Strategies, expected payoff, best replies, and Nash equilibrium. We will restrict our attention to symmetric two-player games from here on. For simplicity, we will label the actions by integers $1, 2, \ldots, n$ rather than $R_1, R_2, \ldots R_n$. Any such game is determined by an $n \times n$ matrix $A = [a_{ij}]$, where a_{ij} is the payoff to any player who takes action i when her opponent takes action j.

Given any such game, a ***mixed strategy*** for either player is a probability distribution over the actions $1, 2, \ldots, n$, which we represent as a column vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ of nonnegative numbers x_i with $\sum_1^n x_i = 1$. The set of all such vectors is the *unit simplex in* $\mathbb{R}^n$; it is denoted Δ_n.

The single action i can then be viewed as the mixed strategy represented by the column vector, denoted $\boldsymbol{e}^i$, that has a 1 in the i^{th} position and zeros elsewhere. Such mixed strategies are called ***pure strategies***. The vectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^n$ are the vertices of the simplex Δ_n, and any strategy $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ equals $\sum_1^n x_i \boldsymbol{e}^i$.

There are two or three ways to think of a mixed strategy $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$.

Interpretation 1. As a randomized strategy for a player who chooses an action at random, with x_i being the probability that she chooses action i. We call such a player an ***x*-player.**

Interpretation 2. In terms of an ***x*-population**. This can be either

2a. A population of $\boldsymbol{x}$-players, or

2b. A population of pure-strategy players, in which, for $i = 1, \ldots, n$, x_i is the proportion of players who always play pure strategy i.

In any of these interpretations, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are mixed strategies, the ***expected payoff*** of an $\boldsymbol{x}$-player, or a random member of an $\boldsymbol{x}$-population, playing against a $\boldsymbol{y}$-player or a random member of a $\boldsymbol{y}$-population, is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i y_j = \boldsymbol{x}^T A \boldsymbol{y} = \boldsymbol{x} \cdot A\boldsymbol{y}.$$

We will regularly use the dot-product notation $\boldsymbol{x} \cdot A\boldsymbol{y}$. Besides avoiding a superscript, the dot helps us remember that $\boldsymbol{x} \cdot A\boldsymbol{y}$ is the payoff to a player using $\boldsymbol{x}$.

We will move freely among the interpretations given above, and we will drop the term "mixed," referring simply to strategies $\boldsymbol{x}$, whether they are mixed or pure. Thus there is no need to distinguish between an i-player and an $\boldsymbol{e}^i$-player. Notice that the expected payoff to an i-player in an $\boldsymbol{x}$-population is $\boldsymbol{e}^i \cdot A\boldsymbol{x}$, which is the i^{th} entry of the column vector $A\boldsymbol{x}$.

Given two strategies $\boldsymbol{x}$ and $\boldsymbol{z}$, we call $\boldsymbol{x}$ a ***best reply*** to strategy $\boldsymbol{z}$ if

$$\boldsymbol{x} \cdot A\boldsymbol{z} \geq \boldsymbol{y} \cdot A\boldsymbol{z} \text{ for all } \boldsymbol{y} \in \Delta_n.$$

A (symmetric) ***Nash equilibrium*** **(NE)** is a strategy $\boldsymbol{x}$ that is a best reply to itself:

$$\boldsymbol{x} \cdot A\boldsymbol{x} \geq \boldsymbol{y} \cdot A\boldsymbol{x} \text{ for all } \boldsymbol{y} \in \Delta_n.$$

If this inequality is strict for all $\boldsymbol{y} \neq \boldsymbol{x}$, then $\boldsymbol{x}$ is a ***strict NE.*** A vertex $\boldsymbol{e}^i$ that is a NE is a ***pure-strategy NE.***

[A general, nonsymmetric NE is a pair $(\boldsymbol{x}, \boldsymbol{y})$ of strategies that are best replies to each other. These certainly may exist in symmetric games; there are even pure-strategy nonsymmetric NE, as shown in (6.2) above. But this concept has less use in evolutionary game theory, so we will not deal with it. ***By "NE" here, we mean symmetric Nash equilibria.***]

In his 1950 PhD dissertation, John Nash proved that ***every finite symmetric two-player game has at least one symmetric NE***. This can be proved using Kakutani's fixed-point theorem. [If $\beta(\boldsymbol{x})$ denotes the set of best replies to $\boldsymbol{x}$, then β is an upper semicontinuous point-to-set map on the compact set Δ_n, and so there is an $\boldsymbol{x}$ with $\boldsymbol{x} \in \beta(\boldsymbol{x})$.]

Following are some important properties of Nash equilibria involving the notion of the ***support*** $S(\boldsymbol{x})$ of a strategy $\boldsymbol{x}$, which is the set of actions i (or the set of vertices $\boldsymbol{e}^i$) for which $x_i > 0$. This might be called the "set of pure strategies in $\boldsymbol{x}$" or the "set of actions available to an $\boldsymbol{x}$-player". Geometrically, it is the set of vertices on the smallest closed face of Δ_n containing $\boldsymbol{x}$. This closed face consists precisely of the convex combinations of the vertices in $S(\boldsymbol{x})$. Notice that the interior of Δ_n is the set of $\boldsymbol{x}$ for which $S(\boldsymbol{x}) = \{1, 2, \ldots, n\}$

The condition $S(\boldsymbol{y}) \subseteq S(\boldsymbol{x})$ means that $y_i = 0$ whenever $x_i = 0$; that is, $\boldsymbol{y}$ contains only actions contained in $\boldsymbol{x}$.

Proposition 7.1. $\boldsymbol{x}$ is a NE if and only if

$$\boldsymbol{e}^i \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x} \text{ for all } i \in S(\boldsymbol{x}) \tag{7.2.1}$$

$$\text{and} \quad \boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x} \text{ for all } i \notin S(\boldsymbol{x}). \tag{7.2.2}$$

That is, if and only if every pure strategy in $\boldsymbol{x}$ is also a best reply to $\boldsymbol{x}$, and the pure strategies not in $\boldsymbol{x}$ are not better replies than $\boldsymbol{x}$ itself.

Proof. If $\boldsymbol{x}$ is a NE, then by definition $\boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x}$ for all i. Since $\boldsymbol{x} = \sum_{i \in S(x)} x_i \boldsymbol{e}^i$, so $\boldsymbol{x} \cdot A\boldsymbol{x} = \sum_{i \in S(x)} x_i \boldsymbol{e}^i \cdot A\boldsymbol{x}$. So $\boldsymbol{x} \cdot A\boldsymbol{x}$ is a weighted average of numbers $\boldsymbol{e}^i \cdot A\boldsymbol{x}$, none of which can exceed it. So they must all equal it.

Conversely, if x satisfies (7.2.1) and (7.2.2), then for any $\boldsymbol{y}$ we have $\boldsymbol{y} \cdot A\boldsymbol{x} = \sum_{i=1}^{n} y_i \, \boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \sum_{i=1}^{n} y_i \, \boldsymbol{x} \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$, so $\boldsymbol{x}$ is a NE. □

Corollary 7.1.1. A strict Nash equilibrium must be one of the pure strategies $\boldsymbol{e}^i$.

Proof: If $\boldsymbol{x}$ is a strict NE but not a pure strategy, then $\boldsymbol{e}^i \cdot A\boldsymbol{x} < \boldsymbol{x} \cdot A\boldsymbol{x}$ for all i, whereas we know $\boldsymbol{e}^i \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$ for $i \in S(\boldsymbol{x})$. □

Proposition 7.2. If $\boldsymbol{y} \in \Delta_n$ and $\boldsymbol{y}$ is a linear combination of best replies to $\boldsymbol{x}$, then $\boldsymbol{y}$ is also a best reply to $\boldsymbol{x}$.

Proof: If $\boldsymbol{x}$ is a NE and $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^k$ are best replies to $\boldsymbol{x}$, then $\boldsymbol{y}^i \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$ for $i = 1, \ldots, k$. If $\boldsymbol{y} = \sum_{j=1}^{k} c_j \boldsymbol{y}^j$ is in Δ_n. then one can check that $\sum_{1}^{k} c_j = 1$, and so

$$\boldsymbol{y} \cdot A\boldsymbol{x} = \sum_{1}^{k} c_j \boldsymbol{y}^j \cdot A\boldsymbol{x} = \sum_{1}^{k} c_j \boldsymbol{x} \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}. \qquad \square$$

Note that although $\sum_{1}^{k} c_j$ must equal 1 for $\boldsymbol{y}$ to be in Δ_n, it is not necessary that $0 \leq c_j \leq 1$ for all j. For example,

$$2 \cdot \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} + 3 \cdot \begin{bmatrix} 0.4 \\ 0.3 \\ 0.3 \end{bmatrix} - 4 \cdot \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}.$$

Corollary 7.2.1. If $\boldsymbol{x}$ is a NE and $\boldsymbol{y}$ is a strategy with $S(\boldsymbol{y}) \subseteq S(\boldsymbol{x}),$ then $\boldsymbol{y} \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$ (i.e., $\boldsymbol{y}$ is also a best reply to $\boldsymbol{x}$).

Proof. Such a strategy is a linear combination of the pure strategies in $\boldsymbol{x}$. □

Corollary 7.2.3. $\boldsymbol{x}$ is a NE if and only if

$$\boldsymbol{e}^i \cdot A\boldsymbol{x} = \max_{y} (\boldsymbol{y} \cdot A\boldsymbol{x}) \text{ for every } i \text{ with } x_i > 0. \tag{7.3}$$

Proof: That (7.3) holds for any NE is immediate from Proposition 7.1. Conversely, if (7.3) holds, then

$$\boldsymbol{x} \cdot A\boldsymbol{x} = \sum_{x_i>0} x_i \boldsymbol{e}^i \cdot A\boldsymbol{x} = \max_{\boldsymbol{y}} (\boldsymbol{y} \cdot A\boldsymbol{x}) \sum_{1}^{n} x_i = \max_{\boldsymbol{y}} (\boldsymbol{y} \cdot A\boldsymbol{x}),$$

so $\boldsymbol{x}$ is a NE.

□

8. Evolutionarily stable strategies. Under Interpretation 2 above, if $\boldsymbol{x}$ is a Nash equilibrium, then we might say that an $\boldsymbol{x}$-population is "resistant to invasion" in the following sense. If it is invaded by a relatively small group of individuals with a different mix $\boldsymbol{y}$, then, because $\boldsymbol{x} \cdot A\boldsymbol{x} \geq \boldsymbol{y} \cdot A\boldsymbol{x}$, the invaders will not do better against the incumbents than the incumbents have been doing against each other.

Clearly, though, what is important is how the invaders and the incumbents will perform in the whole augmented population, not just against the incumbents. Evolutionary stability formalizes this and turns out to be a considerable strengthening of the conditions for a Nash equilibrium.

Imagine an $\boldsymbol{x}$-population – either a group of players all using mixed strategy $\boldsymbol{x}$, or a group of pure-strategy players with mix $\boldsymbol{x}$. Suppose this population is invaded by a small $\boldsymbol{y}$-population, whose size is only ϵ times that of the $\boldsymbol{x}$-population. Then the expected payoff to a $\boldsymbol{z}$-player in the enlarged population, whether $\boldsymbol{z}$ is $\boldsymbol{x}$ or $\boldsymbol{y}$ or one of the pure strategies $\boldsymbol{e}^i$ (or, for that matter, any other strategy that may wander in), is

$$\boldsymbol{z} \cdot A(\epsilon\boldsymbol{y} + (1-\epsilon)\boldsymbol{x}) = \epsilon\boldsymbol{z} \cdot A\boldsymbol{y} + (1-\epsilon)\boldsymbol{z} \cdot A\boldsymbol{x}.$$

We call $\boldsymbol{x}$ an ***evolutionarily stable strategy*** **(ESS)** if, for any strategy $\boldsymbol{y}$, there is an ϵ_y such that as long as $\epsilon < \epsilon_y$ this expected payoff is greater for $\boldsymbol{z} = \boldsymbol{x}$ than for $\boldsymbol{z} = \boldsymbol{y}$. That is,

$$\epsilon\boldsymbol{x} \cdot A\boldsymbol{y} + (1-\epsilon)\boldsymbol{x} \cdot A\boldsymbol{x} > \epsilon\boldsymbol{y} \cdot A\boldsymbol{y} + (1-\epsilon)\boldsymbol{y} \cdot A\boldsymbol{x} \quad \forall \epsilon < \epsilon_y. \tag{8.1}$$

In other words, $\boldsymbol{x}$ is evolutionarily stable if, as long as the invading population is small enough, the incumbents will do better than the invaders in the enlarged population.

If the inequality in (8.1) holds but not strictly, then $\boldsymbol{x}$ is called a ***neutrally stable strategy*** (NSS).

Proposition 8.1. $\boldsymbol{x}$ is an ESS (NSS) if and only if $\boldsymbol{x}$ is a NE that satisfies in addition

$$\text{If } \boldsymbol{y} \neq \boldsymbol{x} \text{ and } \boldsymbol{y} \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}, \text{ then } \boldsymbol{y} \cdot A\boldsymbol{y} < (\leq)\ \boldsymbol{x} \cdot A\boldsymbol{y}. \tag{8.2}$$

Proof. We can rewrite (8.1) as

$$(1-\epsilon)(\boldsymbol{x} \cdot A\boldsymbol{x} - \boldsymbol{y} \cdot A\boldsymbol{x}) + \epsilon(\boldsymbol{x} \cdot A\boldsymbol{y} - \boldsymbol{y} \cdot A\boldsymbol{y}) > 0 \quad \forall \epsilon < \epsilon_y,$$

and the result follows. □

In words, an ESS is a NE $\boldsymbol{x}$ such that, if $\boldsymbol{y}$ is another best reply to $\boldsymbol{x}$, then $\boldsymbol{x}$ is a better reply to $\boldsymbol{y}$ than $\boldsymbol{y}$ itself. This does not look like much of a strengthening, but when we consider the replicator dynamics below, we will see that it is significant.

Corollary 8.1.1. An ESS is a NSS, and a NSS is a NE. □

We will not prove the following useful proposition here; its proof is surprisingly technical.

Proposition 8.2. $\boldsymbol{x}$ is an ESS (NSS) if and only if

$$\boldsymbol{x} \cdot A\boldsymbol{y} > (\geq)\, \boldsymbol{y} \cdot A\boldsymbol{y} \text{ for all } \boldsymbol{y} \neq \boldsymbol{x} \text{ in } N_{\boldsymbol{x}} \tag{8.3}$$

where $N_{\boldsymbol{x}}$ is some neighborhood of $\boldsymbol{x}$, intersected with Δ_n.

The strict version of (8.3) is called ***local superiority***. Proposition 8.2 shows in another light how evolutionary stability strengthens the notion of Nash equilibrium.

Corollary 8.2.1. If $\boldsymbol{x}$ is a NE and every neighborhood of $\boldsymbol{x}$ contains at least one other NE, then $\boldsymbol{x}$ is not an ESS.

> **Proof**. If $\boldsymbol{x}$ were an ESS, then by Proposition 8.2 it would have a neighborhood in which $\boldsymbol{x} \cdot A\boldsymbol{y} > \boldsymbol{y} \cdot A\boldsymbol{y}$ for every $\boldsymbol{y}$. But at least one of these $\boldsymbol{y}$ is a NE, and for this $\boldsymbol{y}$ we have $\boldsymbol{y} \cdot A\boldsymbol{y} \geq \boldsymbol{x} \cdot A\boldsymbol{y}$, a contradiction. □

That is, an ESS must be an isolated NE. However, an NSS need not be, as we see below in Section 13

The following is a restatement of Proposition 8.2.

Corollary 8.2.2. $\boldsymbol{x}$ is an ESS (NSS) iff there exists a positive ϵ_x so that

$$\boldsymbol{\epsilon} \cdot A\boldsymbol{x} + \boldsymbol{\epsilon} \cdot A\boldsymbol{\epsilon} < (\leq)\, 0, \tag{8.4}$$

for all nonzero vectors $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$ with $\sum_1^n \epsilon_i = 0$ and $\sum_1^n \epsilon_i^2 < \epsilon_x$ and $\boldsymbol{x} + \boldsymbol{\epsilon} \in \Delta_n$.

> **Proof.** Vectors $\boldsymbol{y}$ near $\boldsymbol{x}$ in Δ_n are vectors $\boldsymbol{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is as described. It is easy to check that (8.3) is equivalent to (8.4). □

9. Evolutionary game theory: replicator dynamics. Evolutionary stability is clearly designed to confer "invasion resistance" on an $\boldsymbol{x}$-population, but it is a static property, concerning only the relative expected payoffs of various strategies in one play. Now we move to a model of the evolution of a population, in order actually to see what happens when individuals playing different pure strategies come together repeatedly. The model is a system of ordinary differential equations that yield what are called the ***replicator dynamics***. They model evolution through natural selection, with payoffs regarded as evolutionary "fitness points" determining the number of offspring who inherit one's own strategy.

The idea is that we take a symmetric 2-player game with action set $1, 2, \ldots, n$, and we start at time 0 with a population of pure-strategy players that has mix $\boldsymbol{x}(0)$. The i^{th} component $x_i(0)$ is the proportion of the population who are ***i-players***: they always take action i.

Prior to reproducing, individuals play this game many times against randomly-chosen other members of the population. Each individual gains "fitness points" proportional to her total payoff in these games. Then we decree that an individual's number of daughters will be proportional to her number of fitness points, and her daughters will inherit the pure strategy that she plays. As time passes the mix changes, becoming $\boldsymbol{x}(t)$ at time t.

In our motivating examples above, the game is the IPD, and the pure strategies are IPD strategies such as TFT, TF2T, Pavlov, and STFT. The payoffs, as above, are the expected winnings per round. These rounds are infinite but carry a discount factor w, as above. We will assume that all individuals play about the same number of rounds, and that this number is large enough to ensure that all pure strategies encounter each other with the appropriate frequencies. Thus a player's total winnings will be proportional to her expected winnings per round.

If $\boldsymbol{x} = \boldsymbol{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ is the mix at a given time, then the expected payoff of an i-player against a random member of the population is $\boldsymbol{e}^i \cdot A\boldsymbol{x}$, which is just the i^{th} entry of the column vector $A\boldsymbol{x}$.

Let $p_i(t)$ denote the *number* of i-players in the population at time t, and let $p(t) = \sum_{i=1}^{n} p_i(t)$ be the total population size. Then $x_i(t) = \frac{p_i(t)}{p(t)}$ is the proportion of i-players. (The notation can be confusing: $\boldsymbol{x}(t)$ is an $n \times 1$ vector, but $p(t) = \sum_i p_i(t)$ is a scalar.)

> **Discrete replicator dynamics**. (This may be omitted.) Here we think of generations labeled by $t = 0, 1, 2, \ldots$. Each individual reproduces just once in each generation while she is alive. The fundamental assumption is
>
> $$p_i(t+1) = \big(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}(t)\big)p_i(t),\ i = 1, \ldots, n. \qquad (9.1)$$
>
> Here α is a constant common to all members of the population, equal to a baseline

fitness level minus a common death rate. Obviously, α and the entries of A need to be scaled so that the multipliers $(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}(t))$ are not too far from 1, in order to avoid population explosion or implosion.

We get a recursion for the proportions $x_i(t)$ as follows. Summing (9.1) gives

$$\begin{aligned} p(t+1) &= \sum_{i=1}^{n} \big(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}(t)\big) p_i(t) \\ &= \alpha p(t) + \sum_{i=1}^{n} (p(t)x_i(t))\boldsymbol{e}^i \cdot A\boldsymbol{x}(t) \\ &= (\alpha + \boldsymbol{x} \cdot A\boldsymbol{x}(t))p(t). \end{aligned} \tag{9.2}$$

Dividing (9.1) by (9.2), we get

$$x_i(t+1) = \frac{\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}(t)}{\alpha + \boldsymbol{x}(t) \cdot A\boldsymbol{x}(t)} \boldsymbol{x}(t),\ i = 1, \ldots, n. \tag{9.3}$$

We see that a pure strategy that earns its players an above-average number of points per encounter will increase its population share, while one that earns less will decrease. The population will be *stationary*, in that $\boldsymbol{x}(t+1) = \boldsymbol{x}(t)$, when $\boldsymbol{e}^i \cdot A\boldsymbol{x}(t) = \boldsymbol{x}(t) \cdot A\boldsymbol{x}(t)$ for all $i \in S(\boldsymbol{x})$; that is, when all players in the population have the same expected payoff. By Proposition 7.1, the population is stationary at any NE, but not necessarily conversely.

Further analysis of the discrete replicator dynamics is not nearly so easy as for the continuous-time version, which we consider next.

Continuous replicator dynamics. Suppose the population is large enough that we can imagine reproduction taking place continuously in time. The fundamental assumption is

$$\dot{p}_i = \big(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}\big) p_i, \quad i = 1, \ldots, n, \tag{9.4}$$

where $\boldsymbol{x} = \boldsymbol{x}(t)$, $p_i = p_i(t)$, and $\dot{p}_i$ denotes the time derivative $\frac{d}{dt} p_i(t)$. (Remember that $p(t)$ is a scalar while $\boldsymbol{x}(t)$ is a vector.) The constant α is common to all members of the population, incorporating a baseline fitness level minus a death rate that is the same for all players. It may appear that α and the entries of A must be carefully scaled to avoid explosion or implosion of the population, but this turns out not to be crucial. In particular, we will see that α disappears from the equations for the $x_i(t)$.

Summing (9.4) over all i gives

$$\begin{aligned} \dot{p} &= \sum_{i=1}^{n} \big(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}\big) p_i \\ &= \alpha p + \sum_{k=1}^{n} (p x_i) \boldsymbol{e}^i \cdot A\boldsymbol{x} \\ &= (\alpha + \boldsymbol{x} \cdot A\boldsymbol{x}) p. \end{aligned} \tag{9.5}$$

From this and (9.4) we get

$$\begin{aligned} \dot{x}_i = \frac{d}{dt}\frac{p_i}{p} &= \frac{\dot{p}_i p - p_i \dot{p}}{p^2} \\ &= \frac{\dot{p}_i}{p} - x_i \frac{\dot{p}}{p} \\ &= \big(\alpha + \boldsymbol{e}^i \cdot A\boldsymbol{x}\big) x_i - (\alpha + \boldsymbol{x} \cdot A\boldsymbol{x}) x_i. \end{aligned}$$

That is,

$$\boxed{\dot{x}_i = x_i(\boldsymbol{e}^i \cdot A\boldsymbol{x} - \boldsymbol{x} \cdot A\boldsymbol{x}), \;\; i = 1, 2, \ldots, n.} \tag{9.6}$$

The quantity in parentheses on the right is the difference between the expected payoff to an i-player against a random member of the population and the expected payoff to one random player in the population against another. Pure strategies whose players earn above the population average will increase their population shares; those that earn less will shrink in number. If any pure strategy is not represented in the population ($x_i = 0$), it will stay unrepresented.

A mix $\boldsymbol{x}$ is a ***stationary point*** **(SP)** of the replicator dynamics if $\dot{x}_i = 0$ for $i = 1, \ldots, n$. Looking at (9.6), we see that this is equivalent to

$$\boldsymbol{e}^i \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x} \text{ for all } i \text{ with } x_i > 0. \tag{9.7}$$

That is, the expected payoff of an i-player is the same for all i.

Proposition 9.1. All vertices are stationary points.

This is immediate; (9.7) is obviously true if $\boldsymbol{x} = \boldsymbol{e}^i$. □

Proposition 9.2.

a. If a is a positive constant and A' denotes aA (the matrix obtained by multiplying every entry of A by a), then the replicator dynamics (9.6) with A' in place of A have the same orbits, stationary points, and dynamic stability properties, and also the same NE, ESS, and NSS.

b. If c is any real constant and A' denotes $A + cE_i$, where E_i has ones in column i and zeros elsewhere, then (9.6) for A' is identical to (9.6) for A.

Proof (for the static properties NE, ESS, and NSS only). Part a is easy to see. For b, we note that $E_i\boldsymbol{x} = \begin{bmatrix} x_i \\ x_i \\ \vdots \\ x_i \end{bmatrix}$ and so $\boldsymbol{e}^i \cdot E_i\boldsymbol{x} = x_i$ and $\boldsymbol{e}^i \cdot E_i\boldsymbol{x} = x_i \sum_j x_j = x_i$. Thus

$$\begin{aligned} \boldsymbol{e}^i \cdot A'\boldsymbol{x} - \boldsymbol{x} \cdot A'\boldsymbol{x} &= \boldsymbol{e}^i \cdot A\boldsymbol{x} - \boldsymbol{x} \cdot A\boldsymbol{x} + c(\boldsymbol{e}^i \cdot E_i\boldsymbol{x} - \boldsymbol{e}^i \cdot E_i\boldsymbol{x}) \\ &= \boldsymbol{e}^i \cdot A'\boldsymbol{x} - \boldsymbol{x} \cdot A'\boldsymbol{x}. \end{aligned}$$

□

Corollary 9.2.1. Subtracting any row of A from all rows (leaving at least one row of zeros) does not change the replicator dynamics. □

The following proposition assures us that all orbits of the replicator dynamics that start in Δ_n stay in Δ_n.

Proposition 9.3. If $\boldsymbol{x}(0) \in \Delta_n$, then $\boldsymbol{x}(t) \in \Delta_n$ for all $t > 0$.

Proof. Suppose $\boldsymbol{x}(0) \in \Delta_n$. We need to show that $x_i(t) \geq 0$ for all i and $s(t) = 1$ for all t, where $s(t) = \sum_1^n x_i(t)$. Summing (9.6) over all i we see that, as long as $\boldsymbol{x}(t) \in \Delta_n$ we have

$$\begin{aligned} \dot{s}\,(t) &= \sum_1^n \big(x_i(t)\big(\boldsymbol{e}^i \cdot A\boldsymbol{x}(t)\big) - x_i(\boldsymbol{x}(t) \cdot A\boldsymbol{x}(t))\big) \\ &= \boldsymbol{x}(t) \cdot A\boldsymbol{x}(t) - \boldsymbol{x}(t) \cdot A\boldsymbol{x}(t)s(t), \end{aligned}$$

and as long as $\boldsymbol{x}(t) \in \Delta_n$, $s(t) = 1$ and so $\dot{s}\,(t) = 0$.

If any $x_i(t) < 0$ for some t, then there is $\bar{t}$ with $0 < \bar{t} < t$ and $x_i(\bar{t}) = 0$, by the continuity of orbits of ODEs. But then (9.6) shows that $\dot{x}_i(\bar{t}) = 0$, and so we must have $x_i(r) = 0$ for all $r \geq \bar{t}$, contradicting $x_i(t) < 0$. □

This is easily generalized: if $\boldsymbol{x}(t)$ is on any face of Δ_n (i.e. $x_i(t) = 0$ for some given collection of indices i), then $\boldsymbol{x}(r)$ is on that face for all $r \geq t$. This says in particular that if some pure strategies are not in $\boldsymbol{x}(0)$, they cannot appear through evolution. That is, there are no mutations.

Proposition 9.4. A Nash equilibrium is precisely a stationary point for which $\boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x}$ for all i not in $S(\boldsymbol{x})$.

Proof. This is immediate from Proposition 7.1. □

If $\boldsymbol{x}$ is in the interior of Δ_n, then all i are in $S(\boldsymbol{x})$, and so we have

Corollary 9.4.1. If $\boldsymbol{x}$ is in the interior of Δ_n then $\boldsymbol{x}$ is a Nash equilibrium if and only if it satisfies (9.7). That is, all interior stationary points are Nash equilibria. □

Thus we can amplify Corollary 8.1.1 as follows.

Proposition 9.5.

$$\text{ESS} \Rightarrow \text{NSS} \Rightarrow \text{NE} \Rightarrow \text{SP} \quad (\text{and NE} \Leftrightarrow \text{SP for interior points}).$$

10. Dynamic stability of stationary points. Now we are ready to examine the behavior of the replicator dynamics. Because Δ_n is compact and no orbit of the replicator dynamics can go outside Δ_n, all orbits will tend toward either a stationary point or a limit cycle. As it happens, limit cycles do not arise in the examples we consider, so we will focus on stationary points.

A stationary point $\boldsymbol{x}^0$ is called (Lyapunov) ***stable*** if every neighborhood B of $\boldsymbol{x}^0$ contains a neighborhood B^0 of $\boldsymbol{x}^0$ such that if $\boldsymbol{x}(0) \in B^0$ then $\boldsymbol{x}(t) \in B$ for all $t \geq 0$.

$\boldsymbol{x}^0$ is ***asymptotically stable*** if it is Lyapunov stable and in addition there is a neighborhood B^* so that if $\boldsymbol{x}(0) \in B^*$ then $\lim_{t \to \infty} \boldsymbol{x}(t) = \boldsymbol{x}^0$.

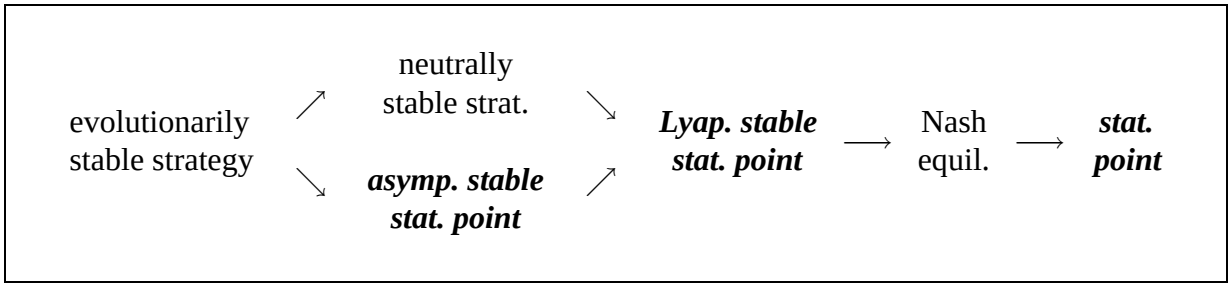
A stationary point that is not Lyapunov stable is called ***unstable***.

Proposition 10.1. Any Lyapunov stable stationary point is a Nash equilibrium.

> **Proof**. Suppose $\boldsymbol{x}$ is a stationary point but not a NE. Then there exists some $\boldsymbol{y}$ with $\boldsymbol{y} \cdot A\boldsymbol{x} > \boldsymbol{x} \cdot A\boldsymbol{x}$. Since $\boldsymbol{x}$ is stationary, we know $\boldsymbol{e^i} \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$ for all i with $x_i > 0$, so there must be some i with $x_i = 0$ and $\boldsymbol{e^i} \cdot A\boldsymbol{x} > \boldsymbol{x} \cdot A\boldsymbol{x}$. Since bilinear functions are continuous, there is some neighborhood U of $\boldsymbol{x}$ and some constant $\delta > 0$ so that $\boldsymbol{e^i} \cdot A\boldsymbol{y} > \boldsymbol{y} \cdot A\boldsymbol{y}$ for $\boldsymbol{y} \in \Delta \cap U$. So for at least this i, if $\boldsymbol{y}^0$ is any point in $\Delta \cap U$ with $y_i > 0$, then for all $\boldsymbol{y} \in \Delta \cap U$ on the orbit starting from $\boldsymbol{y^0}$, $\dot{y}_i > y_i\delta$ and therefore $y_i(t) > e^{\delta t}$. It follows that $\boldsymbol{x}$ cannot be Lyapunov stable. □

The connections among all the important properties of mixed strategies $\boldsymbol{x}$ are summarized in the following chains of implications, extending Proposition 9.5.

Table 10.1



The properties named in italics are defined in terms of the replicator dynamics, and accordingly we will refer to them as "dynamic" stability properties of stationary points. The other properties are defined in terms of relative payoffs in a single play (or IPD round), in attempts to capture evolutionary behavior in a static manner. We will refer to them as "static" stability properties.

We note also that none of the converse implications hold in general, nor does either implication between neutrally stable strategies and asymptotically stable stationary points.

We will not prove these implications here, nor will we give counterexamples to the non-implications.

> The key implications are the two from evolutionary and neutral stability to the two forms of dynamic stability. Lyapunov's direct method is used to prove each of these, and the Lyapunov function is the Kullback-Leibler relative entropy measure $H_x(y) = \sum_i x_i \log(x_i/y_i)$.

The implications in Table 10.1 enable us, at least in some cases, to determine nearly all the dynamic properties of a mix $\boldsymbol{x} \in \Delta_n$ by looking only at the static properties. Only if we discover that a mix $\boldsymbol{x}$ is neutrally stable but not evolutionarily stable would we need to check further to determine whether it is asymptotically stable or only Lyapunov stable.

However, determining the static stability properties and then using Table 10.1 is *not* necessarily the easiest way to investigate the dynamic stability properties. At least in the special cases we will consider below, it is quite easy to identify the stationary points and look at the signs of the derivatives $\dot{x}_i$ at nearby points, to determine their dynamic stability properties without checking their static properties. The point is that knowing the static properties refines our understanding of the various population situations, beyond what we know from the dynamic properties alone.

11. Special case: 2x2 symmetric games. An arbitrary 2×2 symmetric game has a payoff matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. As noted in Corollary 9.1.1, the dynamics will not change if we subtract the first row from both rows. Hence we lose no generality by considering symmetric games with matrices of the form

$$A = \begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix}. \tag{11.1}$$

The mixed strategies, viewed as population mixes – i.e., vectors in Δ_2 – are vectors of the form $\boldsymbol{x} = \begin{bmatrix} x \\ 1-x \end{bmatrix}$, with $0 \leq x \leq 1$, and the replicator dynamics are determined by just one equation in the scalar variable x, namely

$$\dot{x} = x\big(\boldsymbol{e}^1 \cdot A\boldsymbol{x} - \boldsymbol{x} \cdot A\boldsymbol{x}\big). \tag{11.2}$$

But since

$$A\boldsymbol{x} = \begin{bmatrix} 0 \\ ax + b(1-x) \end{bmatrix},$$

(11.2) reads

$$\dot{x} = -x(1-x)(ax + b(1-x)). \tag{11.3}$$

We will first find all stationary points; then we will determine conditions under which each stationary point is a NE, and finally we will determine conditions under which each NE is a NSS or an ESS. As we will see, in this case no point is an NSS without being an ESS, so there will be no open questions about asymptotic stability. Thus we can settle all the questions of dynamic stability of a stationary point in the 2×2 case by examining its static stability properties.

Stationary points. We see immediately from (11.3) that $\dot{x} = 0$ when $x = 1$ and when $x = 0$; that is, the vertices $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ are stationary points. This is always true for any A, by Proposition 9.1.

In the trivial case $a = b = 0$, (11.3) reads $\dot{x} = 0$ and there are essentially no dynamics. It is easy to see that all points are NSS but not ESS in this case. We will assume a and b are not both zero in what follows.

There is a third zero of $\dot{x}$ when $a \neq b$, namely $x = \frac{b}{b-a}$. This will determine a vector in Δ_2 as long as $0 < b < b - a$ or $0 > b > b - a$. This is equivalent to saying that either

$a < 0 < b$ or $a > 0 > b$. In this case we denote the resulting stationary point by $\boldsymbol{x}^0$:

$$\boldsymbol{x}^0 = \frac{1}{b-a}\begin{bmatrix} b \\ -a \end{bmatrix}. \qquad \textbf{(}11.4\textbf{)}$$

Summarizing: $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ are stationary points of the replicator dynamics for (11.1), and if a and b are nonzero with opposite signs, $\boldsymbol{x}^0$ given in (11.4) is also a stationary point.

Nash equilibria. Since all NE are stationary points, we just check the stationary points to see which are NE. By Proposition 9.4, a stationary point $\boldsymbol{x}$ is a NE if and only if, in addition,

$$\boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x} \text{ for all } i \notin S(\boldsymbol{x}). \qquad (11.5)$$

Case 1: $\boldsymbol{x} = \boldsymbol{e}^1$. Here (11.5) says $\boldsymbol{e}^2 \cdot A\boldsymbol{e}^1 \leq \boldsymbol{e}^1 \cdot A\boldsymbol{e}^1$. Note that $\boldsymbol{e}^i \cdot A\boldsymbol{e}^j$ is just the entry in the i^{th} row and j^{th} column of A. So $\boldsymbol{e^1}$ is a NE if and only if $a \leq 0$ (and a strict NE if $a < 0$).

Case 2: $\boldsymbol{x} = \boldsymbol{e}^2$. Now (11.5) says $\boldsymbol{e}^1 \cdot A\boldsymbol{e}^2 \leq \boldsymbol{e}^2 \cdot A\boldsymbol{e}^2$; that is, $0 \leq b$. So $\boldsymbol{e}^2$ is a NE if $b \geq 0$ (and a strict NE if $b > 0$).

Case 3: a and b have opposite signs and $\boldsymbol{x} = \boldsymbol{x}^0$. By Corollary 9.2.1, any interior SP is a NE, so $\boldsymbol{x^0}$ is a NE whenever a and b have opposite signs. ($\boldsymbol{x}^0$ cannot be a strict NE; only vertices can, by Corollary 7.1.1.)

ESS and NSS. All NSS and ESS are NE, so we look at the three cases above to see when the points in question are NSS or ESS.

Case 1: $a \leq 0$ and $\boldsymbol{x} = \boldsymbol{e}^1$. In this case the vectors $\boldsymbol{y}$ in Δ_2 that are in a neighborhood of $\boldsymbol{e}^1$ are vectors $\boldsymbol{y} = \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix}$ for small positive ϵ. For such $\boldsymbol{y}$,

$$A\boldsymbol{y} = \begin{bmatrix} 0 \\ a(1-\epsilon) + b\epsilon \end{bmatrix};$$

so $\boldsymbol{e}^1 \cdot A\boldsymbol{y} = 0$ and $\boldsymbol{y} \cdot A\boldsymbol{y} = \epsilon(a(1-\epsilon) + b\epsilon) = \epsilon[a + \epsilon(b-a)]$.

So $\boldsymbol{e}^1$ is an ESS (NSS) if and only if

$$0 > (\geq)\ \epsilon[a + \epsilon(b-a)] \text{ for small positive } \epsilon. \qquad (11.6)$$

It follows that $\boldsymbol{e}^1$ is an ESS if $a < 0$. If $a = 0$, then (11.6) reads $b\epsilon^2 < (\leq)\ 0$ for small positive ϵ; since we are assuming $b \neq 0$ when $a = 0$, this requires $b < 0$.

Summarizing Case 1: $\boldsymbol{e}^1$ is an ESS if and only if either $a < 0$ (and b is arbitrary), or $a = 0$ and $b < 0$. If $a > 0$, $\boldsymbol{e}^1$ is not even a NE.

Case 2: $b \geq 0$ and $\boldsymbol{x} = \boldsymbol{e}^2$. Neighborhoods of $\boldsymbol{e}^2$ in Δ_2 consist of vectors $\boldsymbol{y} = \begin{bmatrix} \epsilon \\ 1-\epsilon \end{bmatrix}$ for small positive ϵ. For such $\boldsymbol{y}$,

$$A\boldsymbol{y} = \begin{bmatrix} 0 \\ a\epsilon + b(1-\epsilon) \end{bmatrix};$$

so $\boldsymbol{e}^2 \cdot A\boldsymbol{y} = a\epsilon + b(1-\epsilon)$ and $\boldsymbol{y} \cdot A\boldsymbol{y} = (1-\epsilon)(a\epsilon + b(1-\epsilon))$. So $\boldsymbol{e}^2$ is an ESS (NSS) if and only if

$$a\epsilon + b(1-\epsilon) > (\geq)\ (1-\epsilon)(a\epsilon + b(1-\epsilon)) \text{ for small positive } \epsilon. \quad (11.7)$$

This is equivalent to $(a\epsilon + b(1-\epsilon)) > (\geq)\ 0$ for small positive ϵ.

It follows that $\boldsymbol{e}^2$ is an ESS if $b > 0$. If $b = 0$, then (11.7) reads $a\epsilon(1-\epsilon) > (\geq)$ 0 for small positive ϵ. Since we are assuming $a \neq 0$ when $b = 0$, this requires $a > 0$.

Summarizing Case 2: $\boldsymbol{e}^2$ is an ESS if and only if either $b > 0$ (and a is arbitrary), or $b = 0$ and $a > 0$. If $b < 0$, $\boldsymbol{e}^2$ is not even a NE.

Case 3: a and b are nonzero with opposite signs, and $\boldsymbol{x} = \boldsymbol{x}_0 = \frac{1}{b-a}\begin{bmatrix} b \\ -a \end{bmatrix}$. Neighborhoods of $\boldsymbol{x}^0$ in Δ_2 consist of vectors $\boldsymbol{y} = \frac{1}{b-a}\begin{bmatrix} b+\epsilon \\ -a-\epsilon \end{bmatrix}$ for small positive or negative ϵ. For such $\boldsymbol{y}$,

$$A\boldsymbol{y} = A\boldsymbol{x}^0 + \frac{1}{b-a}A\begin{bmatrix} \epsilon \\ -\epsilon \end{bmatrix} = 0 + \frac{\epsilon}{b-a}\begin{bmatrix} 0 \\ a-b \end{bmatrix} = \begin{bmatrix} 0 \\ -\epsilon \end{bmatrix},$$

so $\boldsymbol{x}^0 \cdot A\boldsymbol{y} = \frac{a\epsilon}{b-a}$ and $\boldsymbol{y} \cdot A\boldsymbol{y} = \frac{\epsilon(a+\epsilon)}{b-a}$. So $\boldsymbol{x}^0$ is an ESS (NSS) if and only if $\frac{a\epsilon}{b-a} > (\geq)\ \frac{a\epsilon+\epsilon^2}{b-a}$ for small ϵ; that is,

$$\frac{\epsilon^2}{b-a} < (\leq)\ 0 \quad \text{for all small positive and negative } \epsilon.$$

Since we are assuming a and b are nonzero with opposite signs, this requires $b - a < 0$, i.e. $b < 0 < a$, and implies strict inequality.

Summarizing Case 3: If $b < 0 < a$ then $\boldsymbol{x}^0$ is an ESS. If $a < 0 < b$ then $\boldsymbol{x}^0$ is a NE but not an NSS. In other cases there are no interior SP.

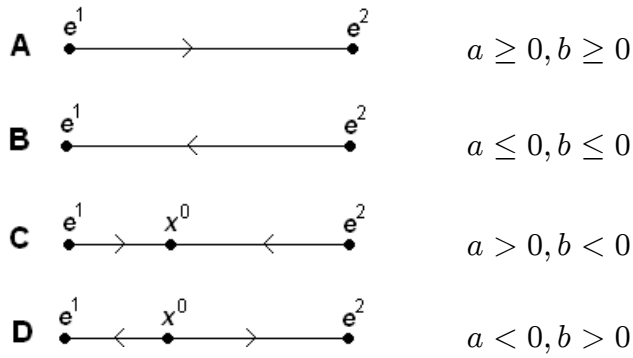
We can put together all the results of this section in the following table. In each case the assertion (SP, NE, NSS, ESS) about each point is the strongest we can make. By Table

10.1, points listed as "NE" or "SP" are unstable stationary points, and those listed as "ESS" are asymptotically stable. Recall that $\boldsymbol{x}^0 = \frac{1}{b-a}\begin{bmatrix} b \\ -a \end{bmatrix}$.

Table 11.1

case	$\boldsymbol{e}^1$	$\boldsymbol{e}^2$	$\boldsymbol{x}^0$	orbit picture
$a = 0,\ b > 0$	NE	ESS	not in Δ_2	**A**
$a = 0,\ b < 0$	ESS	SP	not in Δ_2	**B**
$a > 0,\ b = 0$	SP	ESS	not in Δ_2	**A**
$a < 0, b = 0$	ESS	NE	not in Δ_2	**B**
$a > 0,\ b > 0$	SP	ESS	not in Δ_2	**A**
$a < 0,\ b < 0$	ESS	SP	not in Δ_2	**B**
$a > 0 > b$	SP	SP	ESS	**C**
$a < 0 < b$	ESS	ESS	NE	**D**

With regard to the orbits, the above table shows how the situation depends on the values of a and b. Note that we are excluding the trivial case $a = b = 0$.



None of this is surprising, of course. We note only that the four orbit pictures, which capture the dynamic properties fully, do not reveal the eight different situations with regard to the static properties.

12. Some particular pairs of strategies: Here we look at a few pairs of strategies for the IPD, with various values of the discount factor w.

Example 1: TFT *vs.* STFT. From (5.1) above we get the payoffs

	TFT	STFT
TFT	$\frac{R}{1-w}$	$\frac{S+Tw}{1-w^2}$
STFT	$\frac{T+Sw}{1-w^2}$	$\frac{P}{1-w}$

We multiply the matrix by $1-w^2$ and then subtract the first row from both rows to get

$$A = \begin{bmatrix} R(1+w) & S+Tw \\ T+Sw & P(1+w) \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 \\ (T-R)+w(S-R) & (P-S)+w(P-T) \end{bmatrix}.$$

For the Axelrod numbers $R=3,\ S=0,\ T=5,\ P=1$ this becomes

$$\begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2-3w & 1-4w \end{bmatrix},$$

and we have the following cases, depending on the value of w.

w	case	outcomes
$w \le \frac{1}{4}$	$a>0, b\ge 0$	$\boldsymbol{e}^2$ is ESS; TFT dies out
$\frac{1}{4} < w < \frac{2}{3}$	$b<0<a$	$\boldsymbol{x}^0$ is ESS: $\frac{4w-1}{w+1}$ is stable proportion of TFT
$w \ge \frac{2}{3}$	$a\le 0, b<0$	$\boldsymbol{e}^1$ is ESS; STFT dies out

This is intuitively plausible. Small values of w correspond to shorter games, in which the initial defection of STFT gives it an advantage. Larger values of w nullify this advantage, and it is overcome by the advantage TFT players get by cooperating with each other on every play.

Example 2: TF2T *vs.* STFT. From (5.1) we get

	TF2T	STFT
TF2T	$\frac{R}{1-w}$	$S+\frac{Rw}{1-w}$
STFT	$T+\frac{Rw}{1-w}$	$\frac{P}{1-w}$

and so, multiplying by a positive constant and subtracting the first row from both rows, we get

$$A = \begin{bmatrix} \frac{R}{1-w} & S+\frac{Rw}{1-w} \\ T+\frac{Rw}{1-w} & \frac{P}{1-w} \end{bmatrix} \longrightarrow \begin{bmatrix} R & S+w(R-S) \\ T+w(R-T) & P \end{bmatrix}$$
$$\longrightarrow \begin{bmatrix} 0 & 0 \\ (T-R)(1-w) & (P-S)-w(R-S) \end{bmatrix}.$$

With the Axelrod numbers, this becomes

$$\begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2-2w & 1-3w \end{bmatrix},$$

and we have the following cases, depending on the value of w.

w	case	outcomes
$w \leq \frac{1}{3}$	$a > 0, b \geq 0$	$\boldsymbol{e}^2$ is ESS; TF2T dies out
$w > \frac{1}{3}$	$a > 0, b < 0$	$\boldsymbol{x}^0$ is ESS; $\frac{3w-1}{w+1}$ is stable proportion of TF2T

Example 3: ALC (Always cooperate) *vs.* ALD (Always Defect). Here the payoff table is

	ALC	ALD
ALC	$\frac{R}{1-w}$	$\frac{S}{1-w}$
ALD	$\frac{T}{1-w}$	$\frac{P}{1-w}$

and so the matrix is

$$A = \begin{bmatrix} R & S \\ T & P \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 \\ T-R & P-S \end{bmatrix}.$$

The dynamics do not depend on w, even if the total payoffs do. Because $T > R$ and $P > S$ in any PD, we are in the case $a > 0$, $b > 0$, and so $\boldsymbol{e}^2$ is an ESS and ALC dies out.

13. A class of three-strategy populations. Here we consider a special class of symmetric 3×3 games, with payoff matrices of the form

$$\begin{bmatrix} r & s & t \\ r & s & u \\ v & w & x \end{bmatrix}. \tag{13.1}$$

We will denote the three pure strategies simply by 1, 2, and 3. Notice that if no 3-players are present, neither 1 nor 2 has an advantage; both have the same expected payoff. Any difference in their expected payoffs comes from their encounters with 3-players.

This is motivated by the population dynamics of the three IPD strategies TFT, TF2T, and STFT, whose payoff matrix is given in (5.1) above:

$$A = \begin{bmatrix} \frac{R}{1-w} & \frac{R}{1-w} & \frac{S+Tw}{1-w^2} \\ \frac{R}{1-w} & \frac{R}{1-w} & S + \frac{Rw}{1-w} \\ \frac{T+Sw}{1-w^2} & T + \frac{Rw}{1-w} & \frac{P}{1-w} \end{bmatrix}.$$

The paper of Boyd and Lorberbaum [1987] drew attention to this combination of three strategies. The success of TFT in tournaments like Axelrod's had led to the suggestion that a combination of cooperative strategies like TFT and TF2T would be evolutionarily stable against invasion by any other strategies. Boyd and Lorberbaum showed that TFT by itself is resistant to invasion by STFT, but that a mix of TFT and TF2T can be invaded by a small infusion of STFT, and in the end TFT may die out.

Notice that the payoff matrix for these three strategies specializes (13.1) further in that here $r = s$.

In this section we will study the replicator dynamics of three-strategy populations with payoff matrix of the form (13.1). In the next section we will specialize to the three IPD strategies mentioned above.

If we subtract the first row of (13.1) from all rows, we get a matrix of the form

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a \\ b & c & d \end{bmatrix} \tag{13.2}$$

which produces the same replicator dynamics and the same static equilibrium properties. The equations (9.6) for this A are

$$\dot{x}_i = x_i\big(\boldsymbol{e}^i \cdot A\boldsymbol{x} - \boldsymbol{x} \cdot A\boldsymbol{x}\big), \;\; i = 1, 2, 3.$$

We have

$$A\boldsymbol{x} = \begin{bmatrix} 0 \\ ax_3 \\ bx_1 + cx_2 + dx_3 \end{bmatrix} \quad \text{and} \quad \boldsymbol{x} \cdot A\boldsymbol{x} = x_3(ax_2 + (bx_1 + cx_2 + dx_3)). \tag{13.3}$$

So stationary points must satisfy

$$0 = \dot{x}_1 = -x_1x_3[ax_2 + (bx_1 + cx_2 + dx_3)] \tag{13.4.1}$$
$$0 = \dot{x}_2 = \quad x_2x_3[a - (ax_2 + (bx_1 + cx_2 + dx_3))] \tag{13.4.2}$$
$$0 = \dot{x}_3 = \qquad x_3[\,-ax_2x_3 + (1 - x_3)(bx_1 + cx_2 + dx_3)] \tag{13.4.3}$$

We will look at a few classes of points in Δ_3 and place them on the chain of implications in Table 10.1. We will use the following facts.

$\boldsymbol{x}$ is a SP iff (13.4) holds (or, equivalently, $\boldsymbol{e}^i \cdot A\boldsymbol{x} = \boldsymbol{x} \cdot A\boldsymbol{x}$ for all $i \in S(\boldsymbol{x})$).
$\boldsymbol{x}$ is a NE iff in addition $\boldsymbol{e}^i \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x}$ for all $i \notin S(\boldsymbol{x})$ (Proposition 7.1).
$\boldsymbol{x}$ is an ESS (NSS) iff $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x} + \boldsymbol{\epsilon}) < (\leq)\ 0$ for all sufficiently small vectors $\boldsymbol{\epsilon}$ for which $\boldsymbol{x} + \boldsymbol{\epsilon} \in \Delta_3$ (Corollary 8.2.2).
Any interior SP is a NE.
Non-isolated SP cannot be asymptotically stable, and thus cannot be ESS. (However, they may be Lyapunov stable, and in addition they may be NSS.)

We note also that a SP that is not a NE must be an unstable SP.

There are three classes of points in Δ_3: the vertices, the three open faces (which we refer to as the open "1-2 face" and so on), and the interior.

13a. The vertices. We know that all three vertices are stationary points, by Proposition 9.1. We look at them in turn to see which are NE, NSS, and ESS.

The vertex $\boldsymbol{e}^1$. This point is a NE if and only if $\boldsymbol{e}^i \cdot A\boldsymbol{e}^1 \leq \boldsymbol{e}^1 \cdot A\boldsymbol{e}^1$ for $i = 2$ and $i = 3$. We have $\boldsymbol{e}^1 \cdot A\boldsymbol{e}^1 = \boldsymbol{e}^2 \cdot A\boldsymbol{e}^1 = 0$ and $\boldsymbol{e}^3 \cdot A\boldsymbol{e}^1 = b$, so $\boldsymbol{e}^1$ is a NE if and only if $b \leq 0$.

The vertex $\boldsymbol{e}^1$ cannot be an ESS, because (as we note below) all points on the closed 1-2 face are stationary points, and and ESS must be an isolated stationary point.

$\boldsymbol{e}^1$ is an NSS if and only if $\boldsymbol{\epsilon} \cdot A(\boldsymbol{e}^1 + \boldsymbol{\epsilon}) \leq\ 0$ for all $\boldsymbol{\epsilon} = [-\epsilon_2 - \epsilon_3, \epsilon_2, \epsilon_3]^T$ with ϵ_2 and ϵ_3 nonnegative, not both zero, and sufficiently small. For such an $\boldsymbol{\epsilon}$,

$$\boldsymbol{\epsilon} \cdot A(\boldsymbol{e}^1 + \boldsymbol{\epsilon}) = \epsilon_3[b + \epsilon_2(a - b + c) + \epsilon_3(d - b)].$$

This is certainly nonpositive when $\epsilon_3 = 0$, so $\boldsymbol{e}^1$ is an NSS if and only if $b + \epsilon_2(a - b + c) + \epsilon_3(d - b) \leq 0$ for small nonnegative ϵ_2 and ϵ_3. This is certainly true

if $b < 0$. If $b = 0$, then $\boldsymbol{e}^1$ is an NSS iff $\epsilon_2(a+c) + \epsilon_3 d \leq 0$ for such $\boldsymbol{\epsilon}$; that is, if $a + c \leq 0$ and $d \leq 0$.

Summarizing: $\boldsymbol{e}^1$ is
always a stationary point;
a NE iff $b \leq 0$;
an NSS iff either $b < 0$, or $b = 0$ and $a + c \leq 0$ and $d \leq 0$;
never an ESS.

The vertex $\boldsymbol{e}^2$. A similar argument shows that $\boldsymbol{e}^2$ is
always a stationary point;
a NE iff $a + c \leq 0$;
an NSS iff either $a + c < 0$, or $a + c = 0$ and $b \leq 0$ and $d \leq 0$;
never an ESS.

The vertex $\boldsymbol{e}^3$. This vertex is a NE iff $\boldsymbol{e}^i \cdot A\boldsymbol{e}^3 \leq \boldsymbol{e}^3 \cdot A\boldsymbol{e}^3$ for $i = 1$ and $i = 2$; that is, iff $0 \leq d$ and $a \leq d$.

$\boldsymbol{e}^3$ is an ESS (NSS) iff $\boldsymbol{\epsilon} \cdot A(\boldsymbol{e}^3 + \boldsymbol{\epsilon}) < (\leq)\, 0$ for all $\boldsymbol{\epsilon} = [-\epsilon_2 + \epsilon_3, \epsilon_2, -\epsilon_3]^T$ with ϵ_3 small and positive and $0 \leq \epsilon_2 \leq \epsilon_3$. If we write $\theta\epsilon_3$ for ϵ_2 ($0 \leq \theta \leq 1$), then one can check that for such $\boldsymbol{\epsilon}$,

$$\boldsymbol{\epsilon} \cdot A(\boldsymbol{e}^3 + \boldsymbol{\epsilon}) = \epsilon_3(a\theta - d) + \epsilon_3^2[(-a + b - c)\theta + d - b].$$

So $\boldsymbol{e}^3$ is an NSS iff $a\theta - d \leq 0$ for all $\theta \in [0, 1]$, and, if $a\theta - d = 0$ for some θ (which would have to be $\theta = 0$ or $\theta = 1$), then $(-a + b - c)\theta + d - b \geq 0$ for that θ. The first condition is equivalent to $a - d \leq 0$ and $-d \leq 0$, i.e. to $d \geq 0$ and $d \geq a$. This just says that $\boldsymbol{e}^3$ is a NE. The second condition says in addition that if $d = 0$ then $d - b \geq 0$ (i.e. $b \leq 0$), and if $d = a$ then $-a - c + d \geq 0$ (i.e. $c \leq 0$).

Thus $\boldsymbol{e}^3$ is an NSS iff $\boldsymbol{e}^3$ is a NE and in addition $b \leq 0$ if $d = 0$ and $c \leq 0$ if $d = a$.

Similarly, $\boldsymbol{e}^3$ is an ESS iff

$a\theta - d < 0$ for all $\theta \in [0, 1]$, i.e. $d > 0$ and $d > a$,
or $d = 0 > a$ and $d - b > 0$, i.e. $d = 0, a < 0$, and $b < 0$,
or $d = a > 0$ and $-a + d - c > 0$, i.e. $d = a > 0$ and $c < 0$.

Summarizing: $\boldsymbol{e}^3$ is
always a stationary point;
a NE iff $d \geq 0$ and $d \geq a$,
an NSS iff in addition $d = 0$ implies $b \leq 0$ and $d = a$ implies $c \leq 0$,
an ESS iff in addition $d = 0$ implies $b < 0$ and $d = a$ implies $c < 0$.

One dynamic stability question is left open by the above. $\boldsymbol{e}^3$ will be an NSS without being an ESS in case $b = d = 0 > a$ or $c = d = a > 0$. In such cases we still need to

determine whether $\boldsymbol{e}^3$ is asymptotically stable or merely Lyapunov stable. These cases are exemplified, respectively, by

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a<0 \\ 0 & c & 0 \end{bmatrix} \text{ and } A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a>0 \\ b & a & a \end{bmatrix}.$$

13b. The interior. If x_1, x_2, and x_3 are all required to be nonzero, then the three equations in (13.4) become

$$\begin{aligned} bx_1 + (a+c)x_2 + dx_3 &= 0 \\ bx_1 + (a+c)x_2 + dx_3 &= a \\ bx_1 + (a+c)x_2 + dx_3 &= \frac{bx_1 + cx_2 + dx_3}{x_3}. \end{aligned}$$

So there will be interior stationary points if and only if $a = 0$ and the plane defined by $bx_1 + cx_2 + dx_3 = 0$ passes through the interior of Δ_3. This happens if and only if the vector $[b, c, d]^T$ that is normal to this plane is in neither the closed positive octant nor the closed negative octant.

Thus there will be interior stationary points if and only if $a = 0$ and b, c, and d are all nonzero and not all of the same sign. Moreover, in this case there will be a line segment of interior stationary points, which cannot be ESS, but as noted above must be NE.

Which of these points are NSS? Such a point $\boldsymbol{x}$ is an NSS if and only if $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x} + \boldsymbol{\epsilon}) \leq 0$ for all sufficiently small nonzero $\boldsymbol{\epsilon} = [-\epsilon_2 - \epsilon_3, \epsilon_2, \epsilon_3]^T$ for which $\boldsymbol{x} + \boldsymbol{\epsilon} \in \Delta_3$. Such $\boldsymbol{\epsilon}$ may have positive or negative components. Notice that when $a = 0$ and $bx_1 + cx_2 + dx_3$, we have $A\boldsymbol{x} = [0, 0, 0]^T$. So an interior stationary point $\boldsymbol{x}$ is a NSS if and only if $\boldsymbol{\epsilon} \cdot A\boldsymbol{\epsilon} \leq 0$ for all such $\boldsymbol{\epsilon}$. Thus either all are NSS or none are. When $a = 0$,

$$\boldsymbol{\epsilon} \cdot A\boldsymbol{\epsilon} = (c - b)\epsilon_2\epsilon_3 + (d - b)\epsilon_3^2,$$

and in order that this be nonpositive for all small positive and negative ϵ_2 and ϵ_3, we require $c = b$ and $d - b \leq 0$.

Summarizing:

There are interior stationary points if and only if $a = 0$ and b, c, d are all nonzero and not all of the same sign. In this case there is a line segment of interior stationary points, namely those $\boldsymbol{x} = [x_1, x_2, x_3]^T$ in Δ_3 satisfying $bx_1 + cx_2 + dx_3 = 0$.

These points are all NE. They are all NSS if and only if $b = c \leq d$.

There are no ESS in the interior of Δ_3.

13c. The open 1-3 face. This is the set of vectors $\boldsymbol{x} = [x_1, 0, x_3]^T$ with $0 < x_1 = 1 - x_3 < 1$. For such vectors, (13.4.2) becomes $0 = 0$ and the other two equations in (13.4) both reduce to $b + (d - b)x_3 = 0$. We see that if $b = d = 0$, then all points on the open 1-3 face are SP. If $b = d \neq 0$, there will be none. If $b \neq d$, there will be one SP, with $x_3 = \frac{b}{b-d}$, if $0 < \frac{b}{b-d} < 1$ and none if $\frac{b}{b-d} \leq 0$ or ≥ 1.

Notice that $0 < \frac{b}{b-d} < 1$ if and only if $0 < b < b - d$ or $0 > b > b - d$; that is, if and only if b and d are nonzero with opposite signs.

Which of the SP are NE? $\boldsymbol{x} = [1 - x_3, 0, x_3]^T$ is a NE iff $\boldsymbol{e}^2 \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x}$; that is, if $ax_3 \leq x_3(b + (d - b)x_3)$; that is, $a \leq b + (d - b)x_3$.

In case $b = d = 0$, when all points are stationary, they will all be NE iff $a \leq 0$. In case b and d are nonzero with opposite signs, so that $\boldsymbol{x}^{13} = \left[\frac{-d}{b-d}, 0, \frac{b}{b-d}\right]^T$ is a stationary point, this point will be a NE iff $a \leq 0$. So any SP on the open 1-3 face are NE iff $a \leq 0$.

One of these is an ESS (NSS) if and only if $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x} + \boldsymbol{\epsilon}) <$ ($\leq$) 0 for all sufficiently small nonzero $\boldsymbol{\epsilon} = [-\epsilon_2 - \epsilon_3, \epsilon_2, \epsilon_3]^T$ with $\epsilon_2 \geq 0$ and ϵ_3 positive or negative.

Case 1: If $a \leq 0$ and b and d are nonzero with opposite signs, then $\boldsymbol{x}^{13} = \left[\frac{-d}{b-d}, 0, \frac{b}{b-d}\right]^T$ is a NE. In this case

$$\boldsymbol{\epsilon} \cdot A(\boldsymbol{x}^{13} + \boldsymbol{\epsilon}) = a\epsilon_2 \frac{b}{b-d} + \epsilon_3[(a - b + c)\epsilon_2 + (d - b)\epsilon_3].$$

The first term is a positive multiple of a, so if $a < 0$, then $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x}^{13} + \boldsymbol{\epsilon}) < 0$ for small $\boldsymbol{\epsilon}$, and thus $\boldsymbol{x}^{13}$ is an ESS. If $a = 0$, then $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x}^{13} + \boldsymbol{\epsilon}) = \epsilon_3[(c - b)\epsilon_2 + (d - b)\epsilon_3]$, and unless $b = c = d$ this will take positive values in any neighborhood of $\boldsymbol{x}$, so $\boldsymbol{x}$ is not an NSS.

Case 2: Suppose $a \leq 0 = b = d$, so that all points in the open 1-3 face are NE. None of these can be an ESS because they are not isolated SP and therefore cannot be asymptotically stable. One can check that in this case

$$\boldsymbol{\epsilon} \cdot A(\boldsymbol{x} + \boldsymbol{\epsilon}) = \epsilon_2(ax_3 + (a + c)\epsilon_3),$$

and since $\epsilon_2 \geq 0$, $\boldsymbol{x}$ is a NSS iff

$$ax_3 + (a + c)\epsilon_3 \leq 0$$

for $\boldsymbol{\epsilon}$ as described. Since $x_3 > 0$, we see that if $a < 0$ then this will indeed hold for small $\epsilon_2 \geq 0$ and ϵ_3. But if $a = 0$ then we require $c\epsilon_3 \leq 0$ for all small positive or negative ϵ_3. Thus if $a = 0$ we must also have $c = 0$.

So in case 2 all points on the open 1-3 face are NSS if $a < 0$ or if $a = c = 0$; otherwise none are.

Summarizing:

There are no SP on the open 1-3 face unless either (1) b and d are nonzero with opposite signs or (2) $b = d = 0$.

(1) If b and d are nonzero with opposite signs, then there is exactly one SP on the 1-3 face, namely $\boldsymbol{x}^{13} = \left[\frac{-d}{b-d}, 0, \frac{b}{b-d}\right]^T$. $\boldsymbol{x}^{13}$ is a NE iff in addition $a \leq 0$, and an ESS if $a < 0$. It is never an NSS without being an ESS.

(2) If $b = d = 0$, then all $\boldsymbol{x}$ on the 1-3 face are SP. They are NE iff in addition $a \leq 0$ and NSS iff $a < 0$. They are never ESS.

13d. The open 2-3 face. Vectors $\boldsymbol{x} = [0, 1 - x_3, x_3]^T$ on this face can be handled using the results of 13c above, by converting

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a \\ b & c & d \end{bmatrix} \quad \text{to} \quad A = \begin{bmatrix} 0 & 0 & -a \\ 0 & 0 & 0 \\ b & c & d-a \end{bmatrix},$$

and exchanging strategies 1 and 2 to produce

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -a \\ c & b & d-a \end{bmatrix}.$$

The vector $\boldsymbol{x}$ becomes $[1 - x_3, 0, x_3]$. Translating the above summary merely requires substituting $-a$ for a and $d - a$ for d, and exchanging b and c.

The summary is as follows.

There are no SP on the 2-3 face unless either (1) c and $d - a$ are nonzero with opposite signs or (2) $c = d - a = 0$.

(1) If c and $d - a$ are nonzero with opposite signs, then there is exactly one SP on the 2-3 face, namely $\boldsymbol{x}^{23} = \left[0, \frac{a-d}{a+c-d}, \frac{c}{a+c-d}\right]^T$. $\boldsymbol{x}^{23}$ is a NE iff in addition $a \geq 0$, and an ESS if $a > 0$. It is never a NSS without being an ESS.

(2) If $c = d - a = 0$, then all $\boldsymbol{x}$ on the 2-3 face are SP. They are NE iff in addition $a \geq 0$ and NSS if $a < 0$. They are never ESS.

13e. The open 1-2 face. For $\boldsymbol{x} = [x_1, 1 - x_1, 0]^T$, the equations of (13.4) all reduce to $0 = 0$, and so every point on the open 1-2 face is always a stationary point. (This makes intuitive sense, since neither a 1-player nor a 2-player can distinguish between a 1-player and a 2-player.)

Such an $\boldsymbol{x}$ is a NE iff $\boldsymbol{e}^3 \cdot A\boldsymbol{x} \leq \boldsymbol{x} \cdot A\boldsymbol{x}$, and one can check that this inequality says $bx_1 + c(1 - x_1) \leq 0$; that is, $(c - b)x_1 \geq c$. There are several cases.

If $c > b$, then $(c - b)x_1 \geq c$ iff $x_1 \geq \frac{c}{c-b}$.

If $c > 0 > b$, then $0 < \frac{c}{c-b} < 1$, so $\boldsymbol{x}$ is a NE iff $x_1 \geq \frac{c}{c-b}$.

If $c > b \geq 0$, then $\frac{c}{c-b} \geq 1$, so no such $\boldsymbol{x}$ is a NE.

If $0 \geq c > b$, then $\frac{c}{c-b} \leq 0$, so all such $\boldsymbol{x}$ are NE.
If $c < b$, then $(c-b)x_1 \geq c$ iff $x_1 \leq \frac{c}{c-b}$.
If $c < 0 < b$, then $0 < \frac{c}{c-b} < 1$, so $\boldsymbol{x}$ is a NE iff $x_1 \leq \frac{c}{c-b}$.
If $c < b \leq 0$, then $\frac{c}{c-b} \geq 1$, so all such $\boldsymbol{x}$ are NE.
If $0 \leq c < b$, then $\frac{c}{c-b} \leq 0$, so no such $\boldsymbol{x}$ is a NE.
If $c = b$, then $(c-b)x_1 \geq c$ iff $c \leq 0$.
If $c = b \leq 0$, then all such $\boldsymbol{x}$ are NE.
If $c = b > 0$, then no such $\boldsymbol{x}$ is a NE.

We can rearrange the above into just four cases as follows. For $\boldsymbol{x} = [x_1, 0, 1-x_1]^T$ with $0 < x_1 < 1$:

If $b \leq 0$ and $c \leq 0$, then all such $\boldsymbol{x}$ are NE.
If $b \geq 0$ and $c \geq 0$ (but excluding $b = c = 0$), then no such $\boldsymbol{x}$ is a NE..
If $b < 0 < c$, then all such $\boldsymbol{x}$ with $x_1 \geq \frac{c}{c-b}$ is a NE.
If $c < 0 < b$, then all such $\boldsymbol{x}$ with $x_1 \leq \frac{c}{c-b}$ is a NE.

No NE on the open 1-2 face can be an ESS, because if there are any there will be an open line segment of them, and they will not be isolated NE. An $\boldsymbol{x}$ on the open 1-2 face will be an NSS if and only if $\boldsymbol{\epsilon} \cdot A(\boldsymbol{x} + \boldsymbol{\epsilon}) \leq 0$ for all sufficiently small nonzero $\boldsymbol{\epsilon} = [-\epsilon_2 - \epsilon_3, \epsilon_2, \epsilon_3]^T$ with $\epsilon_3 \geq 0$ and ϵ_2 positive or negative. One can check that this is equivalent to

$$c + (b-c)x_1 + (a-b+c)\epsilon_2 + (d-b)\epsilon_3 \leq 0$$

for small ϵ_2 and ϵ_3 with $\epsilon_2 \geq 0$. We see that $\boldsymbol{x}$ is a NSS iff $c + (b-c)x_1 < 0$; that is, iff $(c-b)x_1 > c$. So we can expand on the four cases above as follows.

All $\boldsymbol{x} = [x_1, 1-x_1, 0]^T$ with $0 < x_1 < 1$ are SP.
If $b \geq 0$ and $c \geq 0$ (but excluding $b = c = 0$), then no such $\boldsymbol{x}$ is a NE.
If $b \leq 0$ and $c \leq 0$, then all such $\boldsymbol{x}$ are NE. All such $\boldsymbol{x}$ are NSS iff in addition $b = c < 0$ or $b < c \leq 0$.
If $b < 0 < c$, then $\boldsymbol{x}$ is a NE iff $x_1 \geq \frac{c}{c-b}$ and a NSS iff in addition $x_1 > \frac{c}{c-b}$.
If $c < 0 < b$, then $\boldsymbol{x}$ is a NE iff $x_1 \leq \frac{c}{c-b}$ and a NSS iff in addition $x_1 < \frac{c}{c-b}$.

As a partial summary of this section, we note the following "watershed" relations among a, b, c, and d.

for $\boldsymbol{e^1}$:
$b = 0$
  $b = 0$, $a + c = 0$, $d = 0$

for $\boldsymbol{e^2}$:
$a + c = 0$
  $a + c = 0$, $b = 0$, $d = 0$

for e^3 :
$d = 0,\ d = a$
$d = 0,\ b = 0$
$d = a,\ c = 0$

for interior :
$a = 0,\ b, c, d$ nonzero and not all same sign

for 1-2 face :
$b = c = 0$

14. The case of TFT, TF2T, and STFT.

Now we apply the above to the particular case of the three strategies TFT (1), TF2T (2), and STFT (3). The matrix A for the replicator dynamics is given by (5.1), but here we begin by simplifying the original PD payoff table $\begin{bmatrix} R & S \\ T & P \end{bmatrix}$. If we subtract S from all entries, and then divide by the positive number $R - S$, we get

$$\begin{bmatrix} 1 & 0 \\ 1+U & V \end{bmatrix} \quad \text{where } U = \tfrac{T-R}{R-S} \text{ and } V = \tfrac{P-S}{R-S}.$$

The conditions $S < P < R < T$ and $R > \frac{1}{2}(S+T)$ are equivalent to $0 < V < 1$ and $0 < U < 1$.

We can interpret U and V as the gain to be had by defecting (i.e. the regret of a cooperator) when the opponent cooperates and defects, respectively, as fractions of $R - S$, the cost of the opponent's defection to a cooperator.

In the special case of the Axelrod numbers $R = 3,\ S = 0,\ T = 5,\ P = 1$, we have $U = \frac{2}{3}$ and $V = \frac{1}{3}$. As we will see, this is a somewhat special situation because $U + V = 1$.

Now we drop the original values and use $R = 1, S = 0,\ T = 1 + U,\ P = V$. Then the matrix given by (5.1),

$$\begin{bmatrix} \frac{R}{1-w} & \frac{R}{1-w} & \frac{S+Tw}{1-w^2} \\ \frac{R}{1-w} & \frac{R}{1-w} & S + \frac{Rw}{1-w} \\ \frac{T+Sw}{1-w^2} & T + \frac{Rw}{1-w} & \frac{P}{1-w} \end{bmatrix}.$$

becomes

$$\begin{bmatrix} \frac{1}{1-w} & \frac{1}{1-w} & \frac{(1+U)w}{1-w^2} \\ \frac{1}{1-w} & \frac{1}{1-w} & \frac{w}{1-w} \\ \frac{1+U}{1-w^2} & \frac{1}{1-w} + U & \frac{V}{1-w} \end{bmatrix}.$$

We multiply all entries by the positive number $1 - w^2$, obtaining

$$\begin{bmatrix} 1+w & 1+w & (1+U)w \\ 1+w & 1+w & w(1+w) \\ 1+U & 1+w+U(1-w^2) & V(1+w) \end{bmatrix},$$

and then subtract the first row from all rows, producing

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -w(U-w) \\ U-w & U(1-w^2) & V-w(U-V+1) \end{bmatrix}.$$

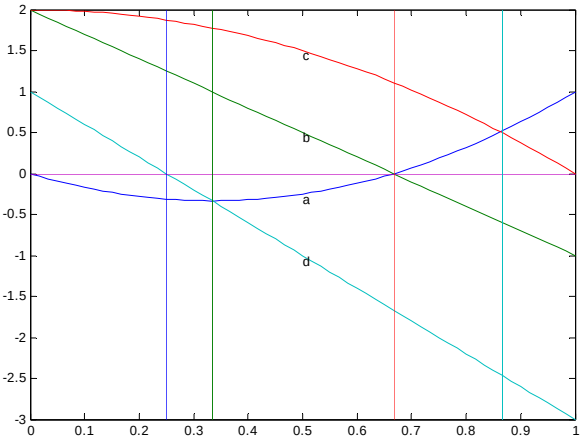
With the Axelrod numbers $U = \frac{2}{3}$ and $V = \frac{1}{3}$, this becomes

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -w\left(\frac{2}{3}-w\right) \\ \frac{2}{3}-w & \frac{2}{3}(1-w^2) & \frac{1}{3}-\frac{4}{3}w \end{bmatrix},$$

and finally we multiply by 3 to get

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & a \\ b & c & d \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -w(2-3w) \\ 2-3w & 2(1-w^2) & 1-4w \end{bmatrix}.$$

The locations of the four parameters with respect to each other and zero, as determined by the value of w, are given in the following graph and table.



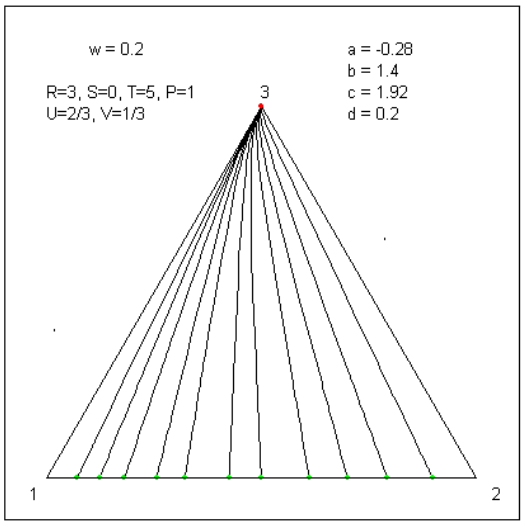
The vertical lines mark the four values of w at which the relative positions of the four parameters and zero change: $w = \frac{1}{4}$, $\frac{1}{3}$, $\frac{2}{3}$, and $\frac{1+\sqrt{11}}{5}$. Thus there are nine separate cases: each of these four values and the five intervals between and around them. However, for the three strategies here, and the particular values of U and V, there are only six different dynamical situations, as shown in the following table.

	parameters		**vertices**			**1-2 face**		**1-3 face**	**2-3 face**	**interior**
#	**values of w**	**a, b, c, d**	**e^1**	**e^2**	**e^3**	$x_1 \leq \frac{b}{b-c}$	**$x_1 > \frac{b}{b-c}$**	$x_1 = \frac{-b}{b-d}$	**$x_2 = \frac{a-d}{a-d+c}$**	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

Following are pictures showing the orbits of the dynamics in each of these cases.

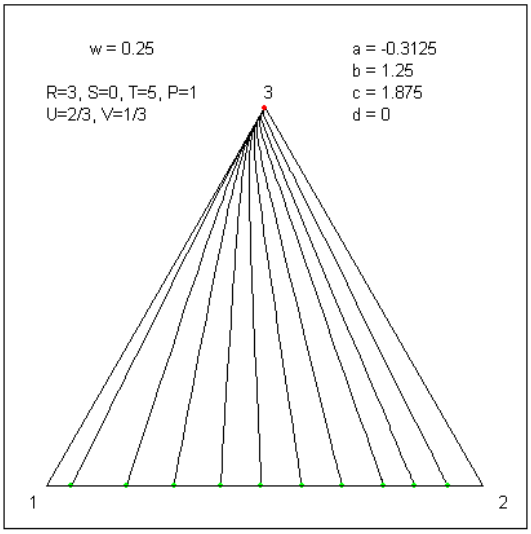
	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a,b,c,d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1=\frac{-b}{b-d}$	$x_2=\frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

Case 1:



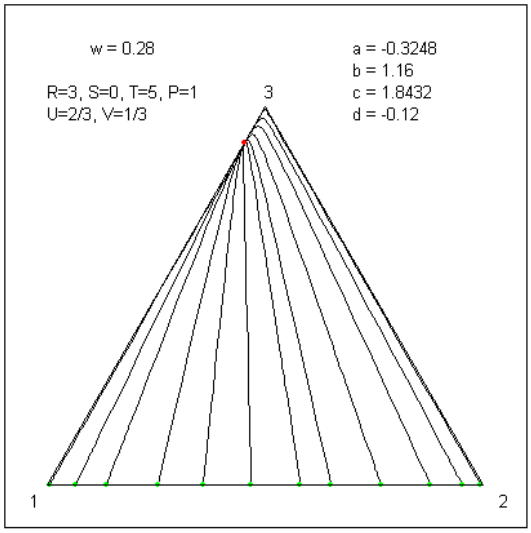
	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a, b, c, d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1 = \frac{-b}{b-d}$	$x_2 = \frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

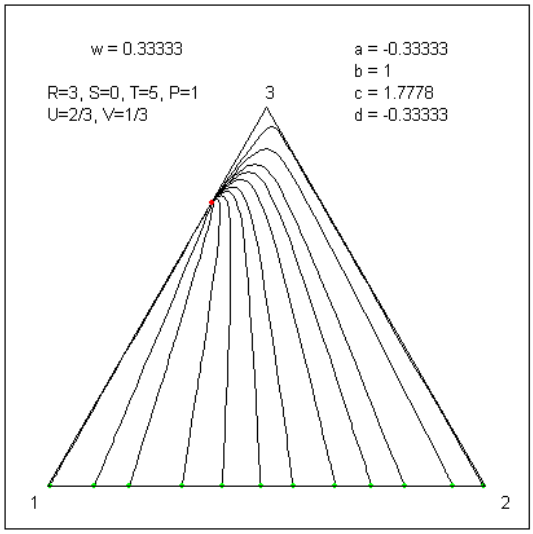
Case 2:



	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a,b,c,d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1=\frac{-b}{b-d}$	$x_2=\frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

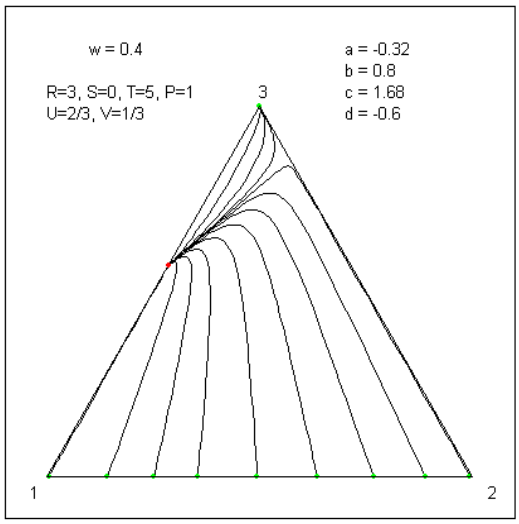
Case 3:

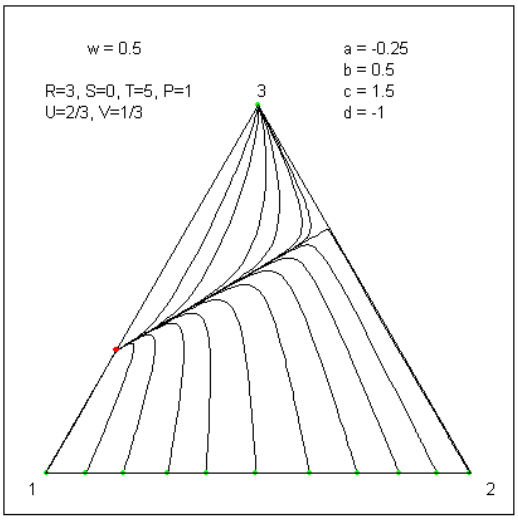


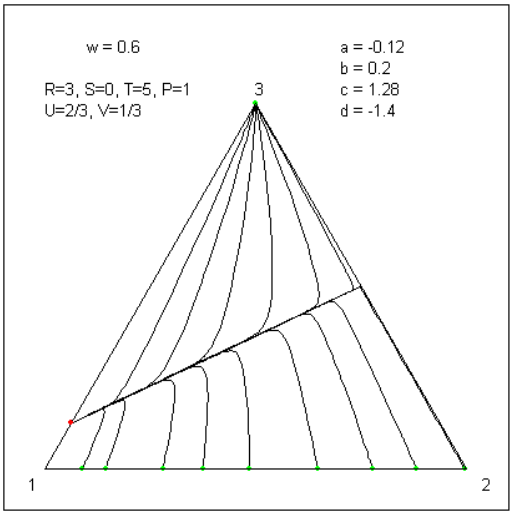


	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a, b, c, d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1 = \frac{-b}{b-d}$	$x_2 = \frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

Case 4:

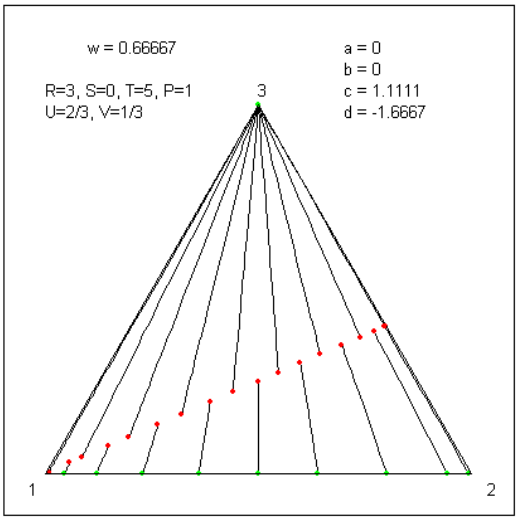






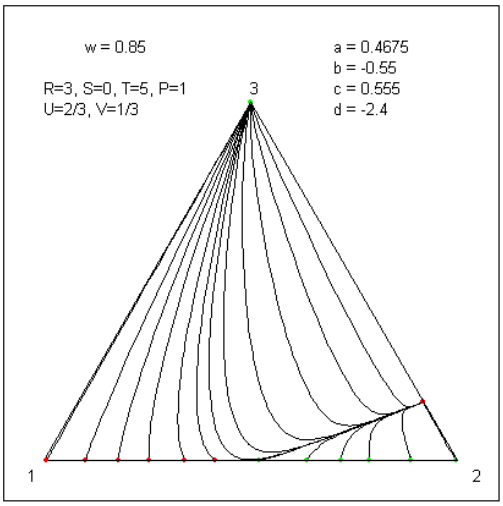
	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a, b, c, d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1=\frac{-b}{b-d}$	$x_2=\frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

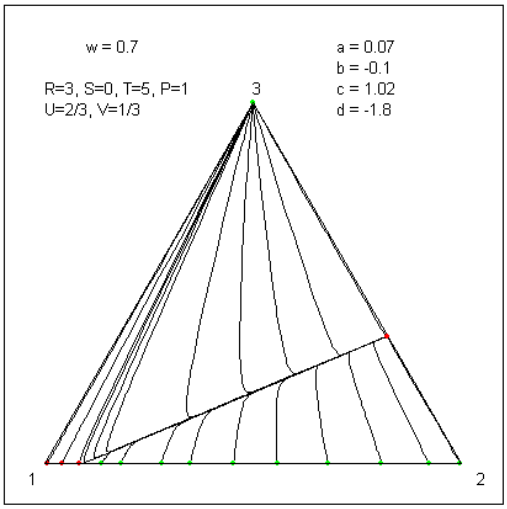
Case 5:

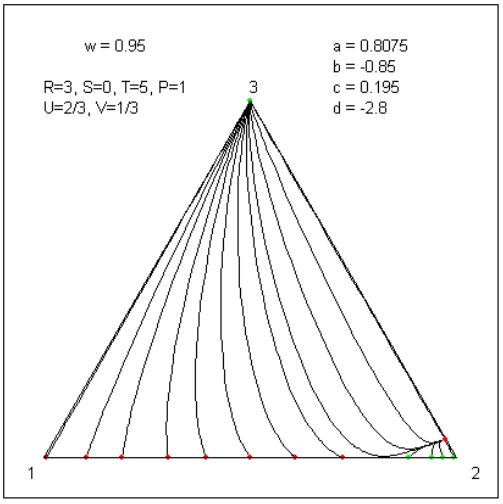


	parameters		vertices			1-2 face		1-3 face	2-3 face	interior
#	values of w	a, b, c, d	e^1	e^2	e^3	$x_1 \leq \frac{b}{b-c}$	$x_1 > \frac{b}{b-c}$	$x_1 = \frac{-b}{b-d}$	$x_2 = \frac{a-d}{a-d+c}$	e^1 to 2-3 face
1	$0 < w < \frac{1}{4}$	$a < 0 < d < b < c$	SP	SP	**ESS**	SP		$\notin \Delta$	$\notin \Delta$	none
2	$w = \frac{1}{4}$	$a < 0 = d < b < c$	SP	SP	**NE**	SP		$\notin \Delta$	$\notin \Delta$	none
3	$\frac{1}{4} < w \leq \frac{1}{3}$	$a < d < 0 < b < c$	SP	SP	SP	SP		**ESS**	$\notin \Delta$	none
4	$\frac{1}{3} < w < \frac{2}{3}$	$d < a < 0 < b < c$	SP	SP	SP	SP		**ESS**	SP	none
5	$w = \frac{2}{3}$	$d < 0 = a = b < c$	**NE**	SP	SP	SP		$\notin \Delta$	**NE**	**NE**
6	$\frac{2}{3} < w < 1$	$d < b < 0 < a, c$	**NSS**	SP	SP	**NSS**	SP	$\notin \Delta$	**ESS**	none

Case 6:







15. References.

Axelrod, Robert: *The Evolution of Cooperation.* Basic Books, 1984.

Axelrod, Robert: *The Complexity of Cooperation.* Princeton University Press, 1997.

Boyd, Robert and Lorberbaum, Jeffrey P.: No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game, *Nature*, vol. 327 (7 May 1987), pp. 58-9.

Hardin, Garrett: The Tragedy of the Commons. *Science*, vol. 162 (1968), pp. 1243-1248.

Hofbauer, J. and Sigmund, K.: *Evolutionary Games and Population Dynamics.* Cambridge University Press, 1998.

Osborne, Martin J.: *An Introduction to Game Theory*. Oxford University Press, 2004 (*an undergraduate-level text*).

Osborne, Martin J.and Rubinstein, Ariel: *A Course in Game Theory*. MIT Press, 1994 (*a slightly higher-level introductory text*).

Poundstone, William: *Prisoner's Dilemma*. Anchor Books, 1993.

Weibull, Jörgen W.: *Evolutionary Game Theory.* MIT Press, 1997.

A website of interest: <http://www.prisoners-dilemma.com>