The Ecology of the Space of 2×2 Social Dilemmas

D.J. Goforth
Department of Mathematics and Computer Science

D.R. Robinson
Department of Economics
Laurentian University, Sudbury ON

May 5, 2004

Abstract

ecology n. 1. The division of biology that treats of the relations between organisms and their environment. 2. *Sociol.* The study of human populations in terms of physical environment, spatial distribution, and cultural characteristics. *Funk & Wagnalls Canadian College Dictionary*, 1986.

Do experiments with a single Prisoner's Dilemma payoff matrix apply to all PDs? Do they ever apply to games that are not PDs? We describe a two-dimensional payoff space of the twelve 2×2 games that allows us to sample payoff matrices in a controlled fashion. Using examples from evolutionary tournament play and from reinforcement learning strategies, we show that resulting populations can vary significantly across the payoff plane and that the boundaries of the ordinal games are not the boundaries for behaviour, evolution or learning.

1 Introduction: When is a PD not a PD? When is a PD a Chicken?

We are accustomed to thinking of games defined in terms of ordinal features. Are ordinally equivalent games behaviourally equivalent? If they are, then studies conducted with a single payoff bi-matrix will provide useful explanations or predictions about the ordinal equivalence class. For example, Prisoner's Dilemma is defined by a set of ordinal constraints on payoffs but the vast majority of simulation experiments are conducted with one payoff matrix popularized by Axelrod in his tournaments[1][2], and presumed to apply to all Prisoner's Dilemmas[1]. We show this implicit extrapolation is

[1] One acknowledgment of the variation of behaviour with payoffs is the quantitative constraint that typically extends the definition of Prisoner's Dilemma to assure that mutual cooperation is the optimal behaviour rather than alternating exploitation.

false.

In this paper we examine how the payoff environment effects the success of strategies competing in the world of 2×2 games. In one experiment we explore the effect of error in a typical evolutionary game replicated with hundreds of payoff matrices. In a second experiment, we observe the behaviour of a learning strategy when it is exposed to scores of different payoff matrices. To set up these experiments we begin by describing a representation on the plane of the twelve symmetric 2×2 games.

2 The environment - a plane of symmetric games

The set of all possible bi-matrices in 2×2 games can only be exhaustively represented in an eight-dimensional space. In an appropriate normalization, it is possible to represent all the *symmetric* games on a 2-D plane. If we restrict attention to symmetric games, only four variables are required. In the terminology commonly used for Prisoner's Dilemma, these are

- C, the *cooperation* payoff when both cooperate
- D, the *defection* payoff when both defect[2]
- T, the *temptation* payoff for defecting when the opponent cooperates
- S, the *sucker* payoff for cooperating when the opponent defects

	Coop	Defect
Coop	C, C	S, T
Defect	T, S	D, D

	Coop	Defect
Coop	3, 3	0, 5
Defect	5, 0	1, 1

At the right are the values used by Axelrod ([1],[2],[3]) that have become the *de facto* standard payoff set for Prisoner's Dilemma.

To produce our representation, we first assume that D is less than C. If D is greater than C, the bi-matrix can be made to conform by swapping rows and columns. We can fix the value of C without loss of generality. We then define the difference between D and C as a unit[3], fixing D also. It is now possible to define a plane with the horizontal dimension representing T, temptation, and the vertical representing S, the sucker payoff.

All symmetric games can be identified with a point in this plane. In Figure 1, the metric is defined using Axelrod's $C = 3$ and $D = 1$. The solid lines on the diagram represent boundaries between the twelve ordinally distinct symmetric games. The payoff matrix used by Axelrod is identified with the point $(5,0)$ in the gray region where S and T satisfy the constraints associated with the three lines: $T > C$, $S < D$ and $S + T < 2C$. By moving this point to different values of S and T, different games are defined[4]. As

[2] C and D are also used to identify the player's choices of Cooperate or Defect. We follow this convention when it is obvious the reference is to a move not a payoff.

[3] If $C = D$, an alternative normalization is required.

[4] To locate any symmetric game (c,d,t,s) on this plane, map c to $C(= 3$ *in this case*$)$, d to $D(= 1)$. Values of S and T are computed by linear transformations:

$$S = C + (s - c)\frac{C - D}{c - d}$$

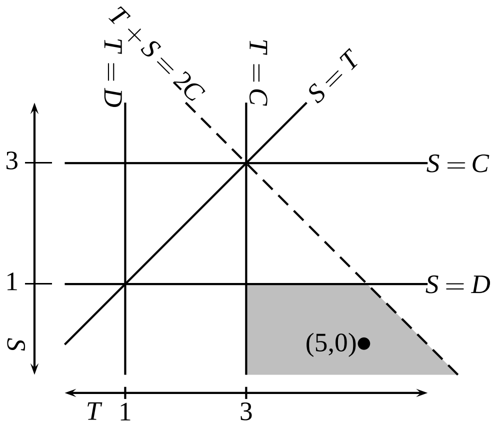

Figure 1: The payoff plane of symmetric 2 × 2 games with Prisoner's Dilemma region in gray

long as the point remains in this region, the game identified with it is a Prisoner's Dilemma[5].

2.1 Mapping the environment: games in their neighbourhoods

Of the symmetric games, six have one equilibrium and are dominance-solvable. The other six have two equilibria. In Figure 2(a), the two distinct sets occupy opposing pairs of quadrants defined by the lines $S = D$ and $T = C$. The dominance-solvable games are gray. The five in the upper left have efficient equilibria while the one in the bottom right is, as we have seen, the Prisoner's Dilemma with its stable but Pareto-dominated equilibrium. The white quadrants are occupied by well-known and problematic games.

The Prisoner's Dilemma together with these multi-equilibria games constitute the *social dilemmas*. In Figure 2(b), the individual regions are labeled with the associated games. Both the Battles-of-the-Sexes and the Coordination Games include a pair of regions which are distinct in terms of the ordering of the S and T payoffs.

The position of Prisoner's Dilemma is central to the other social dilemmas. Any game in the Prisoner's Dilemma region can be transformed into Chicken then into the Battles-of-the-Sexes by increasing the Sucker payoff S. Alternately, a Prisoner's Dilemma game can be transformed to Stag Hunt and the Coordination games by decreasing the Temptation payoff.

Figures 2(c) and (d) show the regions explored in the experiments described in this paper. In section 3, evolutionary experiments are conducted at three scales. In our program, breadth of coverage must be traded against density of sampling so the

and

$$T = C + (t - c)\frac{C - D}{c - d}$$

[5] The dotted line, $S + T < 2C$, is a quantitative constraint on payoffs, restricting Prisoner's Dilemma to the area where mutual cooperation is the optimal strategy.

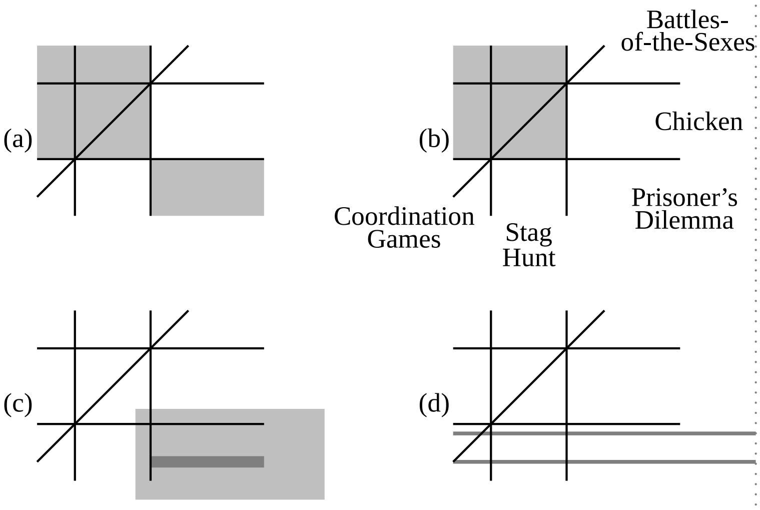

Figure 2: The terrain of the space of 2 × 2 games

- (a) Dominance solvable vs. two equilibria games
- (b) Boring games vs social dilemmas
- (c) Territories explored in evolution experiments ▬ ■ ⬚
- (d) Territories explored in learning experiments ═

sequence represents "zooming" for more detail in a region of interest. In section 4, learning experiments are conducted along two horizontal lines of payoff sets.

3 Ecology of evolution

To explore the effects of actual payoffs on success in playing games, we have used the experimental method introduced in [1], [2], of defining a set of strategies to compete in a round-robin tournament of iterated game play. Depending on success as measured by cumulative payoff, the relative numbers of each strategy in the population are revised and the tournament is replayed. This process is repeated through some number of generations with the intent of reaching an equilibrium population distribution.

The influence of the payoffs can be examined by running the tournament at various points on the payoff plane (i.e., varying the S and T payoffs) and comparing the surviving population profiles.

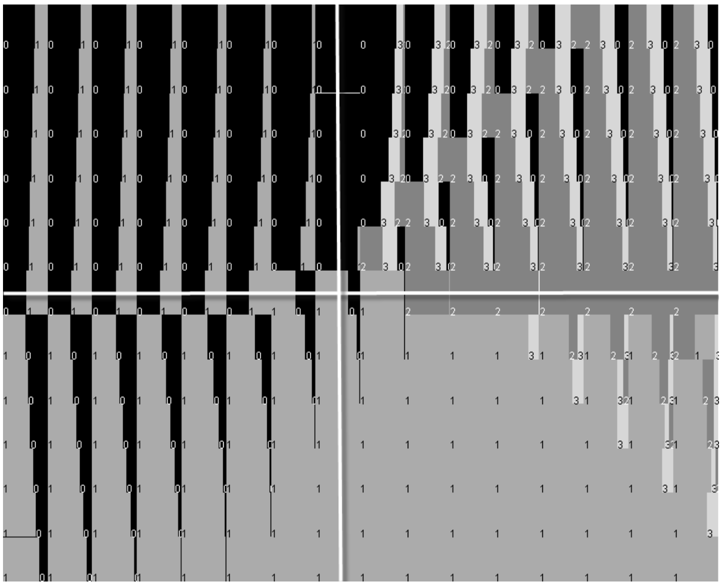

Figure 3: Survival of four populations competing in the payoff plane

All_C	black	always cooperate
All_D	dark gray	always defect
TFT	gray	Tit-for-tat: begin with cooperation, and respond with opponent's last move
aTFT	light gray	anti-Tit-for-tat: begin with defection, and respond with *opposite* of opponent's last move
C_Alt	negative diagonal	begin with cooperation then alternate
D_Alt	positive diagonal	begin with defection then alternate

Table 1: Strategy key for the evolutionary game tournaments

3.1 Four strategies on the payoff plane

As a simple example inspired by [3], consider a small population of four strategies[6], the first four entries in Table 1.

In a tournament with the standard Prisoner's Dilemma payoffs, each game iterated through 200 rounds and the population evolved through 1000 generations. TFT dominated with 99% and *All_C* survived with 1% of the final population. To show the effect of the payoffs on this result, the tournament was rerun with different values for S and T. In our program, this can be done by specifying a range and sampling sequence for each variable, then the results are displayed graphically on a rectangular lattice.

We first look at the big picture. Iterating through odd values of S from -11 to 13

[6] In [3], these strategies played four-round iterated Prisoner's Dilemma and evolved through learning from neighbours in a spatial population. We are not replicating these simulations.

and T between -11 and 19 produces the grid of results in Figure 3. In this composite diagram, each surviving population is shown in a rectangle that is essentially a horizontal stacked bar graph[7], centred at the coordinates (S,T). For example, the fraction of each rectangle that is black indicates the portion of the survivors playing *All_C*. The overall effect is to show how the success of each strategy varies across the payoff plane.

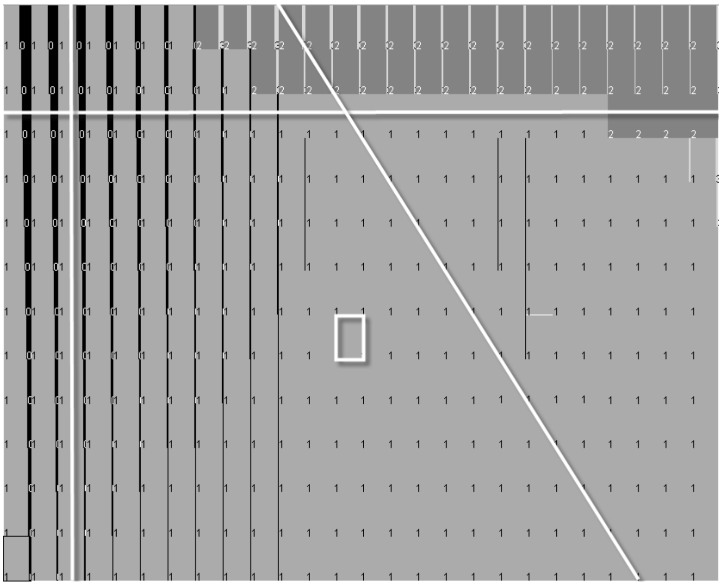

Figure 4: Four populations competing in the Prisoner's Dilemma region

The white lines ($T = C$ and $S = D$) in Figure 3 mark the quadrants defined in Figure 2(a). Starting in the bottom left, TFT dominates the Coordination games but as S increases, *All_C* becomes more and more successful until, near the line $S = D$, it overtakes TFT. In the top left quadrant, *All_C* dominates the games with one equilibrium with TFT surviving. Moving right toward the $T = C$ line, *All_C* continues expanding. In the Battle-of-the-Sexes region, $aTFT$ (light gray) and *All_D* (dark gray) begin to appear. *All_D* becomes the most successful strategy near the horizontal boundary. In the bottom right, TFT reappears and completely dominates most of the quadrant, though $aTFT$ and *All_D* persist in the region beyond Prisoner's Dilemma[8].

To see the details of the population distribution, we "zoom in" producing Figure 4 which shows the results of sampling the smaller range $2.6 \leq T \leq 7.6$ and $-1.0 \leq S \leq 1.4$ with step size 0.2. This is mainly Prisoner's Dilemma territory (bounded by Chicken above and Stag Hunt to the left). Clearly TFT dominates everywhere to all borders and beyond. There is some incursion of *All_C* from the left and $aTFT$ barely survives out beyond the diagonal boundary. At the resolution of this diagram, the 1% of *All_C* does not even show up for the $(5,0)$ rectangle (outlined in white). In this map,

[7] In Excel, this is called a 100% stacked bar graph.

[8] The white and black spots are digits, remnants of strategy identification which may be lost in scale reduction. Zooming on diagrams will restore some detail.

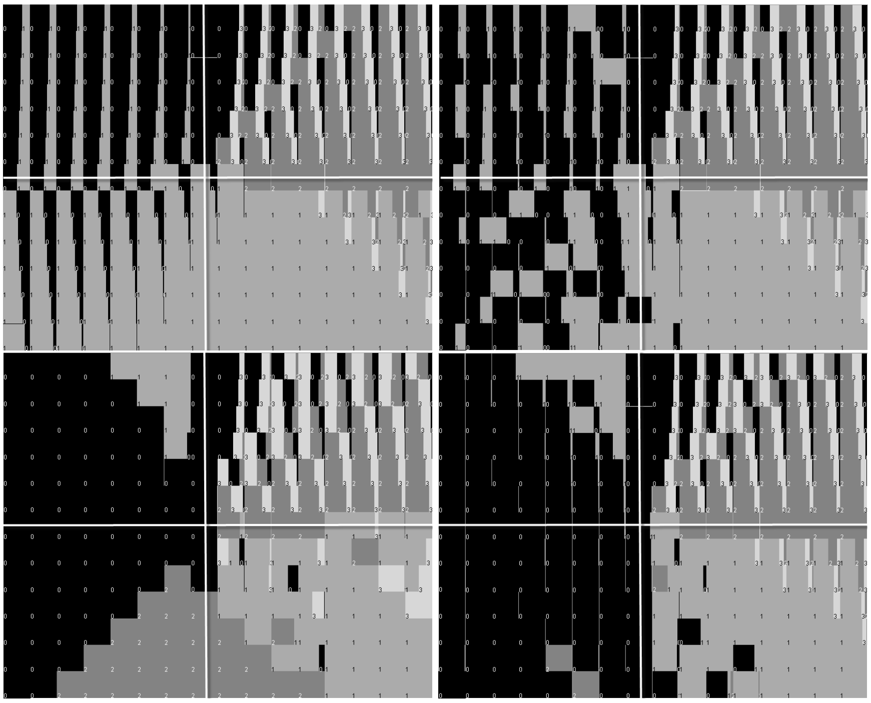

Figure 5: Effect of error across the payoff plane.

Clockwise from top left: error rates 0, $1/10000$, $1/1000$, $1/100$

it is difficult to argue that the results of a classic Prisoner's Dilemma experiment should not be extrapolated to the entire Prisoner's Dilemma region.

3.2 The ecology of errors

When the possibility of error is introduced, the situation changes. We define an error rate m at which an intended move by a player is randomly switched to the opposite move. In Figures 5 and 6, the diagrams of Figure 3 and 4 are repeated at top left. The remaining three panels of each figure show the results of rerunning the experiments with increasing error rates in clockwise order. At the top right, $m = 0.0001$, so approximately one in $10,000$ moves is corrupted. At the bottom right, $m = 0.001$ and at the bottom left, $m = 0.01$.

When $m = 0.00001$, the broad view shows instability in the Coordination games. The apparently minor territorial expansion by All_C into the Prisoner's Dilemma area is revealed in the detailed view as a field of chaotic encounter between All_C and TFT where random factors propel one or the other to dominance, particularly when T is

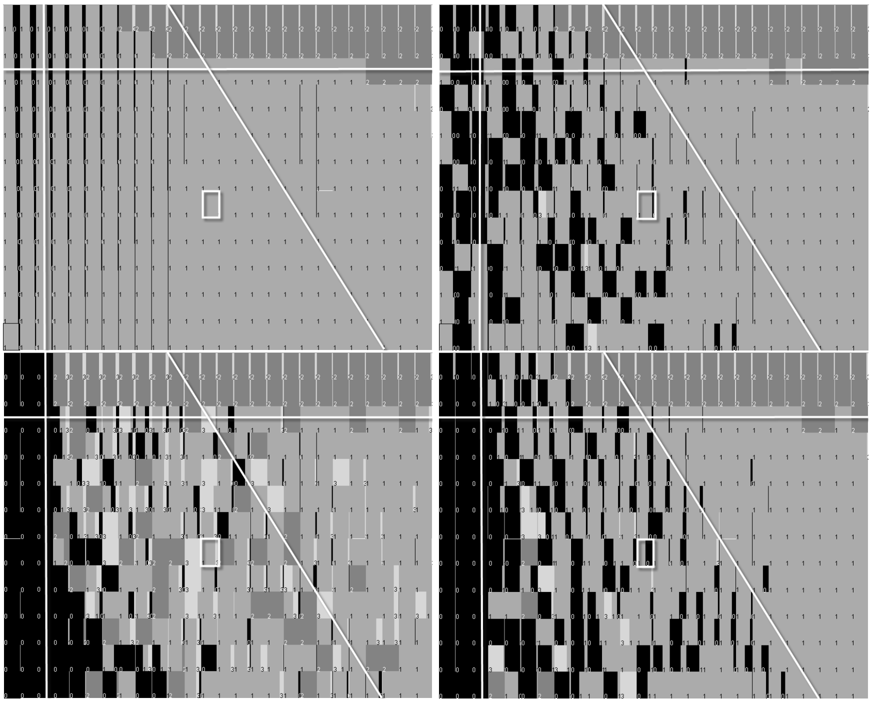

Figure 6: Effect of error in the Prisoner's Dilemma region

closer to C [9]. It appears that unstable evolution is occurring in one area of Prisoner's Dilemma and not in another.

Increasing the error rate to $m = 0.001$(lower right in each figure), causes TFT to become established in the dominance-solvable games of the upper left quadrant while All_C continues to push further across the Prisoner's Dilemma region. However, the situation near the left boundary of Prisoner's Dilemma is very unstable with All_D and $aTFT$ each dominating some populations.

Finally, with an error rate of $m = 0.01$(lower left), All_D dominates a region of Coordination games, Stag Hunt and Prisoner's Dilemma games where S is small and the core area of Prisoner's Dilemma appears to have become extremely unstable. Errors have finally destabilized the region out beyond $T + S = 2C$.

3.3 The ecology of errors with six strategies

To continue this example, we add the final two strategies of Table 1, a pair of *alternators* that cooperate then defect repeatedly. One, C_Alt, begins by cooperating, the other, D_Alt, starts with defection, so they should be symbiotically successful in the region

[9] These results are based on a single run of each tournament to emphasize the instability of the populations with error factored in.

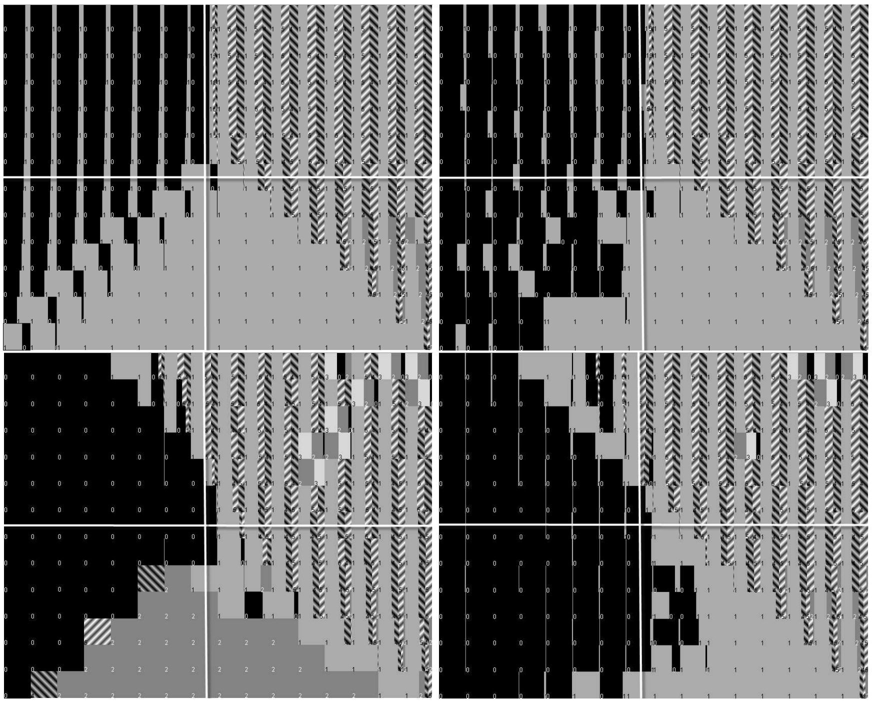

Figure 7: Survival of six strategy populations across the payoff plane.

beyond Prisoner's Dilemma where $S+T > 2C$. In all the figures, C_Alt is striped with negative slope and D_Alt has positive slope.

Figures 7 and 8 retrace the development just described, with the alternators included. The bottom left image in Figure 8 provides the most striking evidence of the sensitivity of populations to small displacements across the payoff plane unrelated to ordinal boundaries. The Prisoner's Dilemma region exhibits two distinct patterns: where T is near to C, the populations appear very unstable with each of the four original strategies coming to dominate in some tournaments; where T approaches the boundary $S+T = 2C$, the populations are stable with combinations of TFT, C_Alt and D_Alt coming to equilibrium in spite of the high error rate.

For more detail, Figure 9 shows population distributions along three cuts through the Prisoner's Dilemma territory from boundary to boundary near $(5,0)$. $3.00 \leq T \leq 6.00$ with stepsize 0.05 and $S \in \{0.00, 0.02, 0.04\}$. The display is altered so the surviving strategies are stacked vertically with heights proportional to evolutionary success. The combination of TFT, D_Alt and C_Alt is stable from the right boundary ($T = 6$) to a point (between 4.1 and 4.5 depending on the row) where the behaviour becomes abruptly erratic. The left side has the characteristic appearance of a chaotic region where randomness can drive the population to any one of many attractors. If an experiment had been conducted only with the classic payoff matrix (white rectangle), the chaotic behaviour would have been missed.

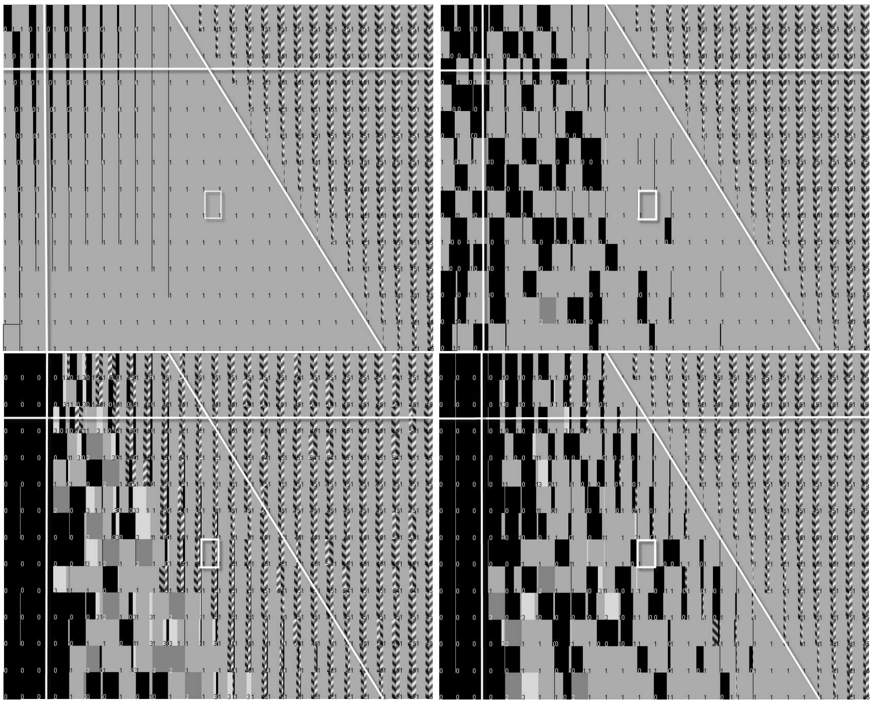

Figure 8: Survival of six strategy populations in the Prisoner's Dilemma region.

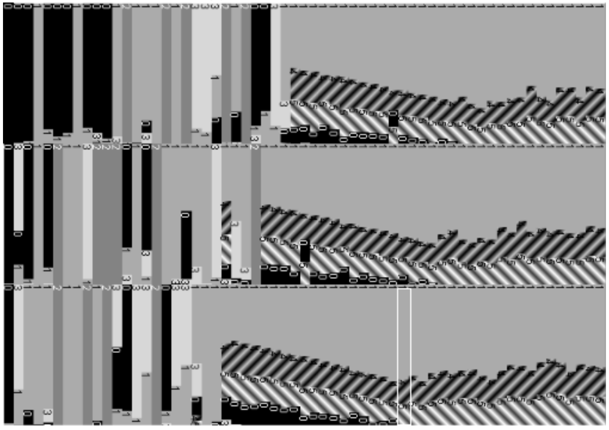

Figure 9: Paths to chaos across the Prisoner's Dilemma region, $m = 0.01$

4 Ecology of Learning

We now consider a strategy that specifically incorporates payoff values into decision-making, a reinforcement learning strategy called Q-Learning. Since the updating of the potential payoff of each possible move is computed from the payoffs in the game, clearly the behaviour of the strategy will be influenced by the actual values of those payoffs.

In this treatment, the learning takes place during a single iterated game. The behaviour changes rather than the population so no evolution is involved. Following an example from [6] and [4], we consider a learning strategy that recognizes it is in one of four possible states defined by the combinations of possible choices of the player and the opponent. These are CC, CD, DC, DD in the notation used previously. The Q-Learning strategy attempts to track the payoff success of making either move, C or D, in each of the states. In other words, there are eight measures of potential payoff.

At each round of iterated play, Q-Learning first updates its estimates of success for the interaction of the last round based on the move made by the opponent and the payoff received. For example, if it chose to defect (D) after a previous round in which both cooperated, CC, then it would update its measure for the combination (CC, D) according to the payoff it got subsequently. Thus it learns by adapting its expectation of future payoff for the move just experienced.

Q-Learning then uses the expectated payoffs to select its next move. The strategy does not simply pick the move with the higher potential for the state it is in. With some probability, it will choose a move randomly instead. This causes Q-Learning to "explore" at least some of the time. As the iteration proceeds and Q-Learning exhausts the new states and moves, the probability of choosing a random move diminishes and Q-Learning comes to trust its recorded measures more and more.

In [6], Sandholm and Crites have examined learning in the Prisoner's Dilemma game by pitting Q-Learning against TFT in a variety of situations. Using the familiar Prisoner's Dilemma payoff matrix, they have designed the learning algorithm to maximize present value of play in an indefinitely continuing iterated game. We concentrate here on their interesting investigation of the sensitivity of the Q-Learning algorithm to future rounds of play. With the future highly discounted, success in the current round dominates the learning and, as would be expected, the algorithm comes to a strategy of always defecting. When future reward is important, the algorithm learns to cooperate. In an intermediate range between these extremes, the algorithm learns to alternate cooperation and defection as best strategy.

Attention to future payoffs is operationalized in a discount factor γ where $0 \leq \gamma < 1$. The algorithm attempts to estimate for each state and action, the value of a series $P_0 + \gamma P_1 + \gamma^2 P_2 + \gamma^3 P_3 + \ldots$ where P_0 is the immediate payoff and remaining P_i are optimal payoffs in future rounds of play. Sandholm and Crites played the learning algorithm against Tit-for-Tat with values of γ between 0.05 and 0.95. The behaviour of the learning algorithm is summarized in Table 2 where DD means that Q-Learning settled in the DD state, indicating it learned to 'always defect.' Similarly, CC is 'always cooperate' and Alt is 'alternately cooperate and defect.' Clearly Tit-for-Tat can collaborate in these interaction patterns. Fang et al. [4] did a similar analysis of Q-Learning

learned strategy	always defect (DD)	alternately cooperate and defect (Alt)	always cooperate (CC)
discount γ	0.05 0.10 0.15 0.20	0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65	0.70 0.75 0.80 0.85 0.90 0.95

Table 2: The strategies learned by Q-Learning against Tit-for-Tat under various future discount factors, γ. From [6]

in the context of the Stag Hunt[10].

In Figure 2.1(d), the two horizontal lines show the sequences of games we have used to replicate the experiments of [6] and [4] with different payoff matrices. On the upper line, the Sucker payoff is 0.75; on the lower, it is 0.0. On each line, $0.0 \leq T \leq 8.0$ with stepsize 0.5. Notice that the sampling includes Coordination games, Stag Hunts, Prisoner's Dilemmas and games beyond the diagonal bound. Games on the boundaries between regions are also included.

The outcome is summarized in Figure 10. T is plotted horizontally. γ ranges vertically; in the ecology of Aesop, the grasshoppers who have no thought of tomorrow leap about at the top while the ants, who care about the future, crawl along below. The original results of Table 2 are outlined in the lower matrix for reference. Notice that one of our values, for $\gamma = 0.25$, does not agree with theirs.

No new behaviour was identified. Q-Learning always comes to one of CC, DD or Alt. However, as the payoffs vary, the proportion of each changes. Between the two matrices of Figure 10, we observe a general pattern that a larger value of the Sucker payoff S, i.e. bringing the Sucker payoff closer to the Defection payoff, causes Q-Learning to adopt the DD strategy only for more extreme discounting of future payoffs.

Starting from the distribution found in [6] with the classic PD payoffs, an *increase* in temptation payoff causes both CC and DD to lose out to Alt. At the $T + S = 2C$ boundary, CC disappears and DD only survives for very shortsighted players. If T *decreases* instead, Alt declines and disappears. For Coordination games, for Stag Hunt and for some PD, Q-Learning switches abruptly between CC and DD at some level of future discount that appears to be fairly independent of the Temptation payoff. Closer examination of γ values near the transition shows that there is not a narrow band of Alt behaviour. Rather, in the boundary region, Q-Learning trials end in either CC or DD, with the proportion of CC varying with γ over a very short interval.

[10] They used these payoffs:

	Coop	Defect
Coop	5, 5	0, 3
Defect	3, 0	1, 1

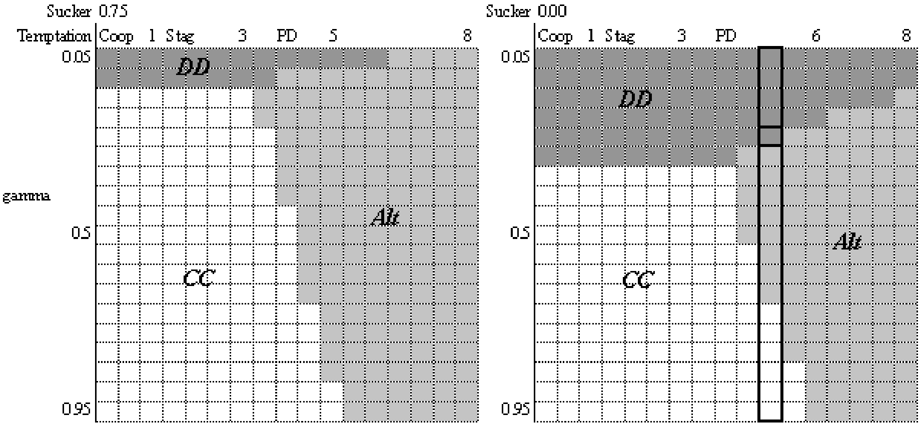

Figure 10: Effect of the discount parameter γ(gamma) on the strategy learned by Q-Learning against Tit-for-Tat

Once again, these data demonstrate that the behaviour of strategies playing the classic Prisoner's Dilemma payoffs cannot be taken as representative of the entire PD region. In fact we can observe that the transition in behaviour here takes place at roughly the same contour through the PD space as the transition from stable alternation to chaotic choice demonstrated in the discussion of evolution with errors in section 3.3.

5 Conclusion

Both the move error experiments of section 3 and the learning example of section 4 identify regions of the PD space with distinct behaviour patterns. On the other hand, the diagrams show that the edges of game regions are not generally locations of abrupt change. Revising payoffs to move among the familiar games such as Prisoner's Dilemma and Stag Hunt is not strongly associated with behavioural transitions in either investigation and this implies an important general result: the discrete order topology of the ordinal games [5] is important to understanding the relationship among 2×2 games but is insufficient for describing and predicting patterns of behaviour.

The two-parameter model of the payoff space of symmetric games shows how the games relate to each other. This model clarifies what the economically significant boundaries are. The diagonal line, $S+T=2C$, really says it all. When $S+T>2C$ (including all Battle-of-the-Sexes games, some Chicken-like games and PD-like games and even some 'boring' games with stable symmetric equilibria[11]), an outcome of one player cooperating and the other defecting, with a payoff of $S+T$, is the most efficient result. In iterated play, the result can also be fair. There has been a tendency to eliminate games with $S+T>2C$ from consideration because the focus has been

[11] For example this game has a dominant strategy equilibrium which is not efficient: $C=3, D=1, T=2, S=5$ Unlike PD, the equilibrium is not Pareto-dominated. Inspect the 'big picture' evolution diagrams to see that some populations can learn to capture this efficiency.

on stimulating true ‘cooperative’ behaviour without collaboration between the players. The alternating choices needed for efficient results seem to imply collaboration but the clear demonstration in [6] that Tit-for-Tat and Q-Learning can achieve alternating behaviour without collaboration should re-open this region of the game space to investigation.

When $S+T \leq 2C$, for all other games, both players should cooperate because $2C$ is the efficient and fair payoff. Three of the Pareto-efficient games are entirely within this half-plane as are the Coordination games and Stag Hunt and Prisoner’s Dilemma.

So, this diagonal line *should* be where changes in behaviour occur. Near the line $S+T = 2C$, the payoffs $S+T$ and $2C$ are similar so it is not surprising that some ’invasions’ take place across the boundary, as strategies engage each other without collusion, especially in conditions of noise and learning.

References

[1] Axelrod, Robert. “Effective Choice in the Prisoner’s Dilemma.” *Journal of Conflict Resolution*, vol 24, no 1, p.3-25, 1980.

[2] Axelrod, Robert. “More Effective Choice in the Prisoner’s Dilemma.” *Journal of Conflict Resolution*, vol 24, no 3, p.379-403, 1980.

[3] Cohen, Michael D., Riolo, Rick L. and Axelrod, Robert. *The Emergence of Social Organization in the Prisoners’ Dilemma: How Context-Preservation and other Factors Promote Cooperation*. Santa Fe Institute Working Paper 99-01-002. 1998.

[4] Fang, Christina, Steven Kimbrough, Annapurna Valluri, Zhiqiang Zheng and Stefano Pace. “On Adaptive Behaviour in the Game of Stag Hunt.” *Group Decisions and Negotiations* 11, 2002, pp449-467.

[5] Robinson, D.R., D.J. Goforth. A topologically-based classification of the 2×2 ordinal games. Presented at the Meetings of the Canadian Economics Association. Carlton University, http://economics.ca/2003/papers/0439.pdf. 2003.

[6] Sandholm, Tuomas and Robert Crites. “Multiagent Reinforcement Learning in the Iterated Prisoner’s Dilemma.” *Biosystems* 37, 1995. pp.147-166.