Iterated Prisoner's Dilemma

Peter Varshavsky

Version of July 14, 2014

Contents

Chapter 1

Introduction

The question that inspired the body of research I focus on in this paper is simple to state: in a world of self-interested individuals, when does it make sense to cooperate? Real life examples abound. Will two profit-minded competing firms engage in price fixing or try to undercut each other? Will athletes take performance-enhancing drugs to gain an edge on their rivals or play fair? Will governments impose tariffs to protect their industries but hurt global output or will they cooperate and remove trade barriers? Will two states locked in an arms race continue building up their weapons arsenal or cooperate and agree to disarmament thus freeing up scarce resources for socially better use.

The question extends beyond human behavior as we observe patterns of cooperation everywhere in nature from reciprocal food exchange in vampire bats [15], to countless examples interspecies cooperation or symbiosis such as the mutualism between cleaner fish and their hosts [1].

These examples share a common structure. Each involves two participants making a choice to cooperate (fix prices, say no to steroids, remove trade restrictions, reduce nuclear weapons stocks, share blood with the less fortunate bat, feed on host's parasites) or defect (undercut the competitor, use steroids, establish trade restrictions, increase weapons stockpiles, hoard the blood, eat the host's mucus instead of the parasites). Mutual cooperation yields a higher advantage to both parties than mutual defection, but if one individual cooperates, while the other defects, the defector gains a large advantage, while the cooperator is left a sucker.

This scenario was originally formalized in 1950 by Merrill Flood and Melvin Dresher at RAND corporation. Soon after Albert W. Tucker offered this illustrative parable that gave the game its name, the Prisoner's Dilemma. In Tucker's tale two prisoners are arrested on a minor charge. The prosecution suspects a more serious crime but does not have evidence to convict unless the prisoners testify against each other. Without the testimony (that is if the prisoners *cooperate (C)* with each other), both would be convicted of the minor charge and receive a small sentence, say one month jail time. However if one prisoner testifies (or *defects (D)* against the other), and the other prisoner keeps mum (cooperates), then the defecting prisoner will have the minor charge dropped, and the cooperating prisoner will receive a large sentence, say five years. Finally if both prisoners rat each other out (or mutually defect), they will both receive large, but lesser sentences, say two years each.

	Mum	Rat
Mum	(-0.1, -0.1)	(-5, 0)
Rat	(0, -5)	(-2, -2)

Because the game is symmetric the ordered pairs in the above matrix are redundant. The usual representation is by a 2×2 matrix of row player's payoffs

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix},$$

with the telling variable names R (*reward* for mutual cooperation), S (*sucker* who cooperates while the opponent defects), T (*temptation* to unilaterally defect), P (*punishment* for mutual defection). The values can vary so long as the ordering $T > R > P > S$ is preserved. We shall often use the following positive values

$$\begin{pmatrix} 3 & 0 \\ 5 & 1 \end{pmatrix}.$$

At first look the proposition may seem pessimistic. When the game is played once by players who have no possibility of future interaction, the only rational move is to defect even though mutual cooperation would lead to a higher payoff for both players. To see this we assume our opponent's move is fixed and analyze our choices. Given that our opponent cooperated, we may choose to cooperate and receive a payoff of $P = 3$, or defect for a payoff of $T = 5$, thus we choose to defect. If our opponent defected, then our choice of cooperation would yield $S = 0$, and defection would pay $P = 1$. Thus regardless of the opponent's choice, it pays for us to defect. In other words if both players choose the strategy of defection, neither can improve her payoff by a unilateral change of strategy. This solution concept, one of the most famous and important in game theory, is called *Nash equilibrium (NE)* in honor of mathematician John Forbes Nash, who showed that in a game with finite sets of players and actions that players can choose from, if players are allowed to randomize their move according to a probability distribution on the space of all possible moves, then at least one Nash equilibrium exists [10]. Game theorists use the term *pure strategies* for the set of all allowed actions, in this case $A = \{C, D\}$. An example of a *mixed strategy* would be deciding to cooperate with some probability p and defect with $1 - p$. If we assume that player X cooperates with probability p and player Y cooperates with probability q, then X's payoff is

$$\begin{aligned} \pi_X &= pqR + p(1-q)S + (1-p)qT + (1-p)(1-q)P \\ &= (R - S - T + P)pq + (S - P)p + (T - P)q + P \end{aligned}$$

differentiating with respect to p,

$$\frac{\partial \pi_X}{\partial p} = q(R - S - T + P) + (S - P) < 0$$

thus independent of Y's strategy, player X will opt to choose the least $p \geq 0$. By symmetry Y would do the same and both players will cooperate with probability $p = q = 0$. So the only Nash equilibrium of the one-shot prisoner's dilemma is mutual defection.

The game becomes more interesting—and less gloomy—with the possibility of repeat interaction, which gives players a chance to reward cooperation and exact vengeance against defectors. When choosing a strategy, one must no longer just maximize the current payoff, but also consider how one's actions will influence the the opponent's response and thus the future payoffs. If the probability of repeat interaction w is large enough, this shadow of the future hangs over every move and encourages players to forego the myopic and stable defection play in search of better outcomes. Because much of the research into iterated prisoner's dilemma is motivated by studying emergence of cooperation, an additional restriction on the payoff values is usually imposed to make sure that repeated mutual cooperation

$$\begin{array}{l} CCCC\ldots \\ CCCC\ldots \end{array}$$

is more profitable than trading defections

$$\begin{array}{l} CDCD\ldots \\ DCDC\ldots \end{array}$$

so in the remainder of the paper we assume

$$R > \frac{T+S}{2}.$$

Whereas in the singleton game the pure strategy space had only two points, the repeated game allows for great complexity and leads to some unexpected results that give insights about many problems in sociology, evolutionary biology, and political science.

The aim of this paper is to survey the history of the Iterated Prisoner's Dilemma focusing on the search for robust strategies that perform well against a wide variety of opponents. I begin with the famous round-robin tournaments conducted by Robert Axelrod in 1980s and present Axelrod's analysis of the simple and elegant winner TIT FOR TAT, which for a time seemed to reign supreme among IPD strategies. Then I explore the insight afforded by borrowing from the language and methods of population biology, Darwinian fitness, and evolutionary stability and present results that unseat TIT FOR TAT and perform well in a wider variety of settings. Finally the second part of the paper focuses on the 2012 discovery of a class of strategies that allow each player to unilaterally force a linear relationship between the players' payoffs leading to opportunities for extortion, dictatorial manipulation of opponent's score, or compliant play. This discovery was followed by a number of papers that explored robustness of the new class of strategies and showed that even with the opportunity to extort, in large populations it may pay to be nice.

Chapter 2

Cooperation

According to Robert Axelrod, there was a rich literature exploring the Prisoner's Dilemma by the late 1970s, but most of it was focused on analyzing situations that shared the game's structure in various social and biological sciences rather than finding the best way to play the game [3, p.28].

Looking to understand how various strategies may perform against each other, he invited fourteen game theorists who specialized in disciplines of psychology, economics, political science, mathematics, and sociology to participate in a computer tournament. Each participant submitted a strategy that described how to play the 200-move iterated Prisoner's Dilemma given the current position and history of prior interaction. In addition Axelrod added a player called RANDOM, which cooperated with probability of one half on every move. Each of the fifteen strategies was set to play each other strategy and itself, and the tournament was repeated five times to get more stable estimates of scores for each pair of players. The score for player i in a match against player j was simply the sum of the payoffs over 200 moves

$$W(I, J) = \sum_{n=1}^{200} \pi_i(x_i^{(n)}, x_j^{(n)})$$

where $x_i^{(n)} \in \{C, D\}$ is i's move at stage n of the game when i follows the strategy I and j follows the strategy J.

The highest possible score $200T = 1000$ could be achieved by an unconditional defector ($AllD$) playing against an unconditional cooperator ($AllC$), who would receive the payoff of $200S = 0$. A pair of $AllC$ players would receive $200R = 600$, and two $AllD$s would earn $200P = 200$.

If players relied on Nash equilibrium, they would all submit the strategy $AllD$, since it is the only Nash equilibrium of the iterated prisoner's dilemma when the number of iterations is finite and known to both opponents. This stems from a backwards induction analysis of possible strategies. On the very last move, without the shadow of the future both players are in effect playing a single-shot game, and thus both defect. Having fully determined the final move, however, they eliminate the shadow of the future from the penultimate move, and thus they again defect. Continuing in this fashion we see that a string of mutual defections, with its accompanying score of 200, is the inevitable result of applying Nash's solution concept. In the case of the tournament, however, players had reason to suspect that there would

be strategies other than *AllD* in the pool which allowed for profitable deviation from the equilibrium.

The simplest submission, entered by Professor Anatol Rapoport, University of Toronto, went on to win the tournament receiving an average score of 504. The strategy, called TIT FOR TAT (TFT), began by cooperating on the first move and on all other moves simply repeated its opponents previous action.

Axelrod circulated the results and solicited entries for the second round of the tournament, this time receiving sixty two entries from professionals and hobbyists. In the first tournament many strategies, knowing that the game only had 200 iterations attempted to use end-of-game tactics to boost their score or to avoid being taken advantage of in the final moves of the game. To avoid this end-of-game opportunism, Axelrod introduced the probability of repeat interaction of $w = 0.99654$, so with probability 0.00346 any iteration could be the last one of the game, and the median match length remained 200 moves.

Allowing for an uncertainty of future interactions fundamentally changes the game. The backwards induction analysis no longer holds, since at no point both players can be sure they shall never meet again, and thus the shadow of the future always looms upon the decision makers. The expected score is now calculated as follows

$$W(I, J) = \sum_{n=0}^{\infty} w^n \pi_i(x_i^{(n)}, x_j^{(n)}).$$

For example the payoff to TFT against the defector *AllD* is

$$W(TFT, AllD) = S + wP + w^2P + w^3P + \cdots = S + \frac{wP}{1-w}.$$

If one or both strategies I, J are stochastic or mixed, then we shall use the same notation $W(I, J)$ to mean the expected value of i's discounted payoff.

Uncertainty of future interaction is mathematically equivalent to discounting time value of money used in economics and finance. This formulation of the game is sometimes called discounted iterated prisoner's dilemma, and it can be shown that with $w > (T-R)/(R-S)$ TIT FOR TAT is now a Nash equilibrium, although it is not unique [3, p.207].

As for the victor of the second tournament, Rapoport submitted TIT FOR TAT for the second time, and once again it won.

The unexpected success of TIT FOR TAT, which can never receive a score higher than its opponent in a single match, was subject of a series of papers and a book *The Evolution of Cooperation* by Axelrod in which he discusses robustness and possible real world examples of the strategy. He suggested that the following four characteristics made TIT FOR TAT successful.

1. *Don't be envious*
   TFT is not trying to beat its opponents, instead it is designed to do well for itself when playing against a wide range of strategies

2. *Don't be first to defect*
   The tournaments showed that strategies that were nice, in the sense that they did not defect unless provoked, did considerably better than strategies that used unprovoked defection

3. *Reciprocate cooperation and defection*
   A strategy that only reciprocates defection, like the grim trigger, gives up the possibility of letting bygones be bygones and returning to mutual cooperation. A strategy that only reciprocates cooperation is easily exploitable by unprovoked defections.

4. *Don't be too clever*
   Rules that constructed complex inferences about their opponents did not do very well likely because the advantages they received from correct inferences did not outweigh the losses due to incorrect ones

Many of these characteristics were shared by other strategies that did well in the tournaments. In fact submissions that tried to get ahead by using unprovoked defections generally did worse than the players that erred on a more forgiving side. In his analysis of the first tournament, Axelrod notes that TIT FOR TAT did not have to win. One strategy that would have done better is the more forgiving TIT FOR TWO TATS (TF2T), which only retaliates after two consecutive defections by its opponent. TF2T, however, is not better than TFT in any absolute way—it was submitted in the second tournament by John Maynard Smith and only finished in 24th place. The poor repeat performance of TF2T was due to it being exploited by strategies that tested how much defection they could get away with. It is a good illustration of the principle that there can not be a single best strategy independent of the population of other strategies: although these testing strategies spelled doom for TF2T, they themselves did not do well in the end [3, p.47].

To further study the robustness of TIT FOR TAT, Axelrod turned to the approach of evolutionary or ecological stability developed by John Maynard Smith in which the average payoff that a strategy receives in a given population determines its reproductive fitness. Each generation of strategies plays a round-robin discounted IPD tournament similar to the second tournament described above, and relative frequency of strategies in the next generation is determined by relative frequencies and average payoffs in the current generation. The chronology of this ecological computational experiment is interesting. In the early generations strategies that are too nice are exploited by meaner strategies, which leads to proliferation of mean strategies and near extinction of too forgiving ones. The mean strategies are usually not very good at playing against other mean strategies, and as their numbers increase while the prey becomes sparse, their fitness suffers, and they eventually lose prominence. TFT, which receives the payoff of $R = 3$ against itself and other nice strategies, and which does not allow exploitation by mean strategies, once again proved to be very robust in this ecological setting and came out on top in every simulation [3, p.53].

The idea that to remain ecologically or evolutionarily fit a strategy needs to perform well against opponents, but also against itself was first formalized by John Maynard Smith in his 1973 paper "The Logic of Animal Conflict" [9]. He defined a strategy J to be an *evolutionarily stable strategy (ESS)* if a large population employing J could not be invaded by a mutant playing some other strategy I, or more precisely if $W(I, J)$ is the payoff to I in

a contest with J, and $W(J,J)$ is the payoff to J playing against itself, then

$$\begin{cases} W(I,J) < W(J,J) \\ \text{ or } \\ W(I,J) = W(J,J) \text{ and } W(J,I) > W(I,I) \end{cases} \tag{2.1}$$

In his 1981 paper "The Emergence of Cooperation among Egoists" [4], Axelrod defined a relaxation of ESS. He called a strategy *collectively stable (CS)* if $W(I,J) \leq W(J,J)$ for any mutant invading strategy I in a population of strategies J. Whereas ESS is has more stringent requirements than the Nash equilibrium, CS is equivalent to a symmetric weak Nash equilibrium restated in the language of evolutionary game theory. This modification allowed him to provide a characterization theorem for this class of strategies. The underlying idea for the characterization is that the incumbent can prevent the newcomer from invading if no matter what the newcomer does, the incumbent can keep his score sufficiently low. Axelrod introduces the definition: J has a *secure position* over I on move n if no matter what I does from move n onwards, $W(I,J) \leq W(J,J)$, assuming that B defects from move n onwards. If $W_n(I,J)$ denotes I's discounted cumulative score in the moves before move n, then we can say that J has a secure position over I on move n if

$$W_n(I,J) + w^{n-1}P/(1-w) \leq W(J,J)$$

or equivalently

$$W_n(I,J) \leq W(B,B) - w^{n-1}P/(1-w) \tag{2.2}$$

thus the characterization theorem, which is a version of the folk theorem of repeated games, "embodies the advice that if you want to employ a collectively stable strategy, you should only cooperate when you can afford an exploitation by the other side and still retain your secure position" [4, p.313].

Theorem 2.1. ***The characterization theorem.*** *J is a collectively stable strategy if and only if J defects on move n whenever the other player's cumulative score so far is too great, specifically when*

$$W_n(I,J) > W(J,J) - w^{n-1}(T + wP/(1-w)). \tag{2.3}$$

Proof. The converse implication uses induction. Assume J is a strategy that defects as required by 2.3. To have a secure position over I on move $n = 1$, J is required to defect on every move, that is J is the $ALLD$ strategy, and $W(I, ALLD) \leq W(ALLD, ALLD)$ for any strategy I, so J has a secure position over I on the first move.

If J has a secure position over I on move n, we need to show that it will maintain its secure position on move $n+1$. First, if J defects on move n, I gets at most P, so

$$W_{n+1}(I,J) \leq W_n(I,J) + w^{n-1}P$$

Using 2.2, we get

$$\begin{aligned} W_{n+1}(I,J) &\leq W(J,J) - w^{n-1}P/(1-w) + w^{n-1}P \\ &\leq W(J,J) - w^n P/(1-w) \end{aligned}$$

J is only allowed to cooperate on move n if J can afford to be exploited by I, that is

$$W_n(I,J) \leq W(J,J) - w^{n-1}(T + wP/(1-w)).$$

Since I can get at most T on the n'th move,

$$\begin{aligned} W_{n+1}(I,J) &\leq W(J,J) - w^{n-1}(T + wP/(1-w)) + w^{n-1}T \\ &= W(J,J) - w^n P/(1-w) \end{aligned}$$

thus if J follows the prescription of the theorem, it will preserve its secure position over I and will be collectively stable.

The forward implication is proved by contradiction. Suppose that J is collectively stable, and there is an I and an n such that J does not defect on move n when

$$W_n(I,J) > W(J,J) - w^{n-1}(T + wP/(1-w))$$

or, equivalently, when

$$W_n(I,J) + w^{n-1}(T + wP/(1-w)) > W(B,B). \tag{2.4}$$

Now define I' to be the same as I on the first $n-1$ moves, and from move n onward. Since J cooperates on move n, I' gets T, and at least P on the following moves. So

$$W(I',J) \geq W_n(I,J) + w^{n-1}(T + wP/(1-w)).$$

combined with 2.4 this yields $W(I',J) > W(J,J)$, so I' invades J, and hence J is not collectively stable. □

Clustering and emergence of cooperation

Axelrod's tournaments and Maynard Smith's ideas of evolutionary stability have given a rather positive answer on rationality of cooperation among self-interested individuals, but it seems at odds with the stability of *AllD*. If we can assume that at some point cooperation did not exist, then the stability of *AllD* implies that cooperation would never emerge from individual mutations, since cooperating individuals would not be able to invade a population of defectors.

This pessimistic prediction changes if we allow cooperators to arrive in groups to the population of defectors. Such clusters are likely to develop among close kin or individuals occupying the same or neighboring territories. Richard Dawkins, in his best seller *The Selfish Gene* [6], describes how such kin-based cooperation or altruism can be modeled and understood if we focus not on survival and procreation of individuals in a population, but of units of genetic information.

To model invasion by a cluster, we still assume that the majority of the population is using the strategy J, but now the newcomers I are arriving in sufficient numbers and sufficiently near each other so that the probability that I interacts with I is $p \in (0,1)$, and the probability that I interacts with J is $1-p$. Then the expected score for a newcomer is

$pW(I,I) + (1-p)W(I,J)$, and for the incumbent $W(J,J)$. So the p-cluster of newcomers I will invade J if $pW(I,I) + (1-p)W(I,J) > W(J,J)$, or, solving for p,

$$p > \frac{W(J,J) - W(I,J)}{W(I,I) - W(I,J)}.$$

We can now calculate the size of the cluster necessary for TIT FOR TAT to invade *AllD*. Given our usual values of T, R, P, S, and $w = 0.9$, we get $p > 1/21$, that is if TFT newcomers have just a five percent chance of interactions with other TFT players, they can successfully invade a world of noncooperative players. As w increases, the minimum value of p necessary for TIT FOR TAT to invade is reduced even lower [4].

Invasion in clusters provides a believable mechanism for cooperation to take hold in a population, and prompts two natural questions. Are there strategies that are most efficient in invading populations of defectors in a sense of requiring the smallest p-cluster to gain a foothold, and would clustering also help mean strategies to invade populations of nice strategies. Axelrod provided the encouraging answers to both. He called a strategy *maximally discriminating* if it will eventually cooperate even if the other has never cooperated yet, and once it cooperates will never cooperate again with *AllD* but will always cooperate with another player using the same strategy as it uses, and proved that the strategies which can invade *AllD* in a cluster with the smallest value of p are those which are maximally discriminating, such as TIT FOR TAT [4, Theorem 7]. He also found that if a nice strategy can resist invasion by a single individual playing an alternate strategy, then it cannot be invaded by any cluster of individuals playing the same strategy [4, Theorem 8].

More on ESS, collective stability, and Nash equilibrium

So far this paper described several solution concepts all based on the idea that in an equilibrium it does not pay to unilaterally change behavior: Nash equilibrium, ESS, and collective stability. We denote the set of strategies that are in Nash equilibrium with themselves, also called a *symmetric Nash equilibrium,* Δ^{NE}, the set of ESS strategies Δ^{ESS}, and the set of collectively stable strategies Δ^{CS}. It follows immediately from definitions that $\Delta^{NE} = \Delta^{CS}$. The relationship between Δ^{ESS} and Δ^{CS} is more interesting. Our definition of ESS (Equation 2.1) is equivalent to

$$\Delta^{ESS} = \{I \in \Delta^{NE} : W(J,J) < W(I,J) \ \ \forall J \in \beta^*(I), \ \ J \neq I\}$$

where $\beta^*(I) = \{J \in \Delta : W(J,I) \geq W(J',I) \ \ \forall J' \in \Delta\}$ is the set of best replies to a strategy I. In particular $\Delta^{ESS} \subset \Delta^{CS}$, so ESS is a more demanding requirement than collective stability. Axelrod motivated his relaxation of ESS to CS as a way to simplify the proofs of his propositions, and claimed that, except for the characterization theorem (Theorem 2.1), they hold for ESS strategies as well. However in a 1987 paper "No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game" [5], R. Boyd and J.P. Lorberbaum showed that neither TFT nor any other strategy whose behavior on a given move is fully determined by the history of prior moves is evolutionarily stable. In his arguments Axelrod considers invasion of a pure strategy population by a single newcomer strategy, and he only considers ecologic dynamics, that is the dynamics of relative frequencies of a set of given strategies

not altered by mutation. If the possibility of mutation is introduced, however, Boyd and Lorberbaum prove that collectively stable strategies can succumb to invasion by pairs of other strategies.

Theorem 2.2. *No strategy whose behavior during interaction n is uniquely determined by the history of the game up to that point is evolutionarily stable if*

$$w > \min\left[\frac{T-R}{T-P}, \frac{P-S}{R-S}\right]$$

Proof. ***(The proof is quoted almost verbatim from [5])*** Let J_{CS} be a collectively stable strategy and let I be a distinct strategy that behaves exactly the same way with J_{CS} on each interaction with J_{CS} as J_{CS} does against itself. This implies that $W(I, J_{CS}) = W(J_{CS}, J_{CS}) = W(J_{CS}, I) = W(I, I)$. If a third strategy K exists in the population, J_{CS} can be invaded by I if $W(I, K) > W(J_{CS}, K)$. Because J_{CS} and I are distinct, there must be some history h such that J_{CS} and I react differently to K for the first time on move n. There are two possibilities. First, suppose J_{CS} defects and I cooperates on move n. Let L be the strategy that generates the history h in response to both J_{CS} and I, cooperates on move n and then defects forever in response to defection by J_{CS} and cooperates forever in response to cooperation by I on move n. J_{CS} can be invaded by I whenever

$$W(I, L) - W(J_{CS}, L) \geq w^{n-1}(R - T) + w^n \frac{R-P}{1-w} > 0$$

or $w > (T-R)/(T-P)$. Next, let M be a strategy which behaves exactly like L for the first $n-1$ moves, defects on move n, and then defects forever in response to J_{CS}'s defection and cooperates forever in response to cooperation by I on move n. In this case J_{CS} can be invaded by I whenever $w > (P-S)/(R-S)$. The second possibility is that J_{CS} cooperates and that I defects on move n. A similar argument shows that J_{CS} can be invaded by I for any value of w. □

Boyd and Lorberbaum note that this theorem agrees with Axelrod's insight that no strategy is strictly a best strategy regardless of the remaining population. They are careful to not imply that their theorem disproves the robustness of nice, provokable and forgiving strategies, but instead shows the existence of scenarios where such strategies, robust as they are, can still be invaded and displaced. They present the following scenario in which TIT FOR TAT is invaded by a combination of the more forgiving TIT FOR TWO TATS (TF2T), and SUSPICIOUS TIT FOR TAT (STFT) which defects on the first interaction and then plays TIT FOR TAT for the remainder of the game. In their example STFT is assumed to be maintained in the population by one-way mutation or phenotypic variation. Because $W(TF2T, TFT) = W(TFT, TFT)$ and since the forgiving $TF2T$ can induce $STFT$ to cooperate, while TFT and $STFT$ end up locked in a string of alternating defections, $W(TF2T, STFT) > W(TFT, STFT)$ when w is large enough, so TF2T can invade a predominantly TFT population if it also contains some STFT individuals.

Win-Stay, Lose-Shift

One of the main features that makes TFT so attractive for studying emergence of cooperation is its simplicity. It does not take a great leap of faith to imagine reciprocity-based behaviors

emerge in human society or the animal kingdom. TFT does not require a long memory or sophisticated optimization calculations, it does not require that the individuals are from the same tribe or species, and that the payoffs are equal or even measured in the same units for the players involved. Whether or not it precisely described the real evolution of cooperation (which it probably did not) is not relevant. Its success lies in giving a simple and constructive proof that a simple system that starts as a population of defectors can develop into a cooperating system. But how unique is TFT in those regards?

In 1993 Martin Nowak and Karl Sigmund published a paper in Nature [11] that proposed that another simple and realistic rule called WIN-STAY, LOSE-SHIFT (WSLS) could outperform TIT FOR TAT. Dubbed PAVLOV for its reflex-like response to the payoffs, the rule repeats its previous move if it was rewarded by R or T points, but changes from C to D or from D to C if it was punished by receiving only S or P. The discovery of WSLS's dominance was accidental. Nowak and Sigmund were studying the performance of various memory-one strategies (strategies that only use the previous move in determining their next move) in populations where errors were possible, that is there was a nonzero probability that TFT would mistakenly cooperate in response to a defection or vice versa as if due to to unsteadiness of players' trembling hands. If we restrict our attention to memory one strategies, iterated prisoner's dilemma can be thought of as a Markov chain with states $\{CC, CD, DC, DD\}$, then a strategy can be described as a probability of that the player plays C given that the game is in any of the above states $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Thus TFT is described by $\mathbf{p}^{TFT} = (1, 0, 1, 0)$, and WSLS corresponds to $\mathbf{p}^{WSLS} = (1, 0, 0, 1)$. GENEROUS TIT FOR TAT (GTFT) with $\mathbf{p}^{GTFT} = (1, \gamma, 1, \gamma)$ is a variant of TFT that cooperates with some probability γ after the opponent's defection (often $\gamma = .3$). Possibility of mistakes can be modeled by changing all unit probabilities in these representations to $1 - \epsilon$ and all null probabilities to ϵ, where ϵ is the probability of error.

When the possibility of errors is added to the model pairs of TFT run the risk of getting stuck in lengthy periods of alternating defection following an unintended defection by one of the players. The generous morph GTFT has the advantage of being able to break these vicious cycles sooner than TFT, which increases GTFT's fitness in the new model. In fact Nowak and Sigmund were expecting a version of GTFT to come be the winner of their simulations, but in the very long run WSLS showed to be more fit, likely owing to its ability to take advantage of unconditional or very forgiving cooperators, while playing well against itself. If, due to its own mistaken defection, WSLS discovers that it can get away with defecting unilaterally, it will continue doing so until it either switches to C due to another error or the opponent's retaliation. Unlike a pair of TFTs, a pair of WSLS players do not get stuck in long loops of alternating defections following the mistake by one of the players—it only takes two moves to return to the state of mutual cooperation.

Nowak and Sigmund started their simulation with a population of random strategies $\mathbf{p}^{\mathrm{RAND}} = (0.5, 0.5, 0.5, 0.5)$. To allow for errors they restricted $0.001 < p_i < 0.999$ for all i for all strategies. On average every 100 generations they introduced a randomly generated mutant into the mix. The simulations were consistent with Axelrod's predictions that TFT mutants can invade predominantly ALLD populations, and with Boyd and Lorberbaum's findings that TFT can be invaded by mixtures of generous and suspicious newcomers. While the chronology of the simulations varied greatly, Nowak and Sigmund point to several interesting features that were consistent. Most of the time was spent in states of various equilibria

lasting from tens of thousands to millions of generations. The transitions between equilibria were rare but fast, often taking just a few generations. Cooperations was observed in only 27.5% of the runs after $t = 10^4$ generations, but in 90% at $t = 10^7$. PAVLOV dominated 10% of the runs at $t = 10^4$ and 82.5% at $t = 10^7$. The periods spent in cooperating equilibria increased in length with time, but the threat of breakdown and reversion to defecting states never completely abated.

Conclusion of Part 1

This concludes the first part of the paper. It focused on emergence of cooperation mainly through the lens of Robert Axelrod tournaments. It compared the solution concepts of ESS by Maynard Smith and collective stability by Axelrod. Axelrod's approach was mainly ecological, meaning mutations were not allowed in the model. Once we enable evolutionary processes by allowing mutations, some of Axelrod's findings are no longer valid. His insight, however, can still be useful.

It is interesting that the strategies that do best in these tournaments are some of the simplest rules: TFT and WIN-STAY-LOSE-SHIFT. This simplicity makes it believable that similar processes could have actually been present in evolution of species, ecosystems, and behaviors.

Chapter 3

Manipulation

Much of the innovation in the history of IPD came from various refinements of the Nash equilibrium. Reframing the equilibrium in terms of ecology led to Axelrod's idea of collective stability and his analysis of the success of TIT FOR TAT. An evolutionary refinement of by Maynard Smith allowed the introduction of mutations and a stability robust to them. It exposed a weakness of TFT—it could be invaded by a pair of strategies. To account for possibility of error or miscommunication, Richard Selten [12] proposed the trembling hand perfect equilibrium, and simulations showed that a simple strategy WSLS performed well in this error-prone environment. Not all successful strategies have the simple elegance of TFT and WSLS, however. In 2010 Iliopoulos, Hintze, and Adami showed that stability of strategies depended on the mutation rate, and analyzed a strategy they called GENERAL COOPERATOR (GC) with $\mathbf{q}_{\text{GC}} = (0.935, 0.229, 0.266, 0.42)$, which was the evolutionary fixed point at a rate of mutation [8].

In 2012 a new class of strategies was born out of yet another approach to solving the dilemma. If the IPD is restricted to memory-one strategies, which can only use the previous state of the game in each decision, it can be thought of as a Markov process with states CC, CD, DC, DD. Since the strategies can be described as vectors of probabilities the player will cooperate given each of the four states of the game

$$\mathbf{p} = (p_1, p_2, p_3, p_4)$$
$$\mathbf{q} = (q_1, q_2, q_3, q_4)$$

So the previous state of CC corresponds with next move probabilities of cooperation p_1 and q_1, CD corresponds to p_2 and q_3, DC to p_3 and q_2, and DD to p_4 and q_4. Then the Markov transition probabilities are

$$\begin{bmatrix} p_1q_1 & p_1(1-q_1) & (1-p_1)q_1 & (1-p_1)(1-q_1) \\ p_2q_3 & p_2(1-q_3) & (1-p_2)q_3 & (1-p_2)(1-q_3) \\ p_3q_2 & p_3(1-q_2) & (1-p_3)q_2 & (1-p_3)(1-q_2) \\ p_4q_4 & p_4(1-q_4) & (1-p_4)q_4 & (1-p_4)(1-q_4) \end{bmatrix}.$$

If $\mathbf{v}$ is the stationary distribution of the above chain, then we can define the long term

average payoffs to be

$$S_X = \mathbf{v} \cdot \mathbf{S}_X = \mathbf{v} \cdot (\mathrm{R}, \mathrm{S}, \mathrm{T}, \mathrm{P})$$
$$S_Y = \mathbf{v} \cdot \mathbf{S}_Y = \mathbf{v} \cdot (\mathrm{R}, \mathrm{T}, \mathrm{S}, \mathrm{P}).$$

Press and Dyson's discovery was that the dot product of the stationary distribution $\mathbf{v}$ with an arbitrary 4-vector $\mathbf{f}$ is given by the following determinant

$$\mathbf{v} \cdot \mathbf{f} \equiv D(\mathbf{p}, \mathbf{q}, \mathbf{f}) = \det \begin{bmatrix} -1 + p_1 q_1 & -1 + p_1 & -1 + q_1 & f_1 \\ p_2 q_3 & -1 + p_2 & -q_3 & f_2 \\ p_3 q_2 & p_3 & -1 + q_2 & f_3 \\ p_4 q_4 & p_4 & q_4 & f_4 \end{bmatrix},$$

in which the second column is $\tilde{\mathbf{p}} \equiv (-1 + p_1, -1 + p_2, p_3, p_4)$ is entirely controlled by the first player, and the third column $\tilde{\mathbf{q}} \equiv (-1 + q_1, -p_3, -1 + q_2, q_4)$ is entirely controlled by the second player. This allows each player to unilaterally make the determinant vanish by setting their column to be a scalar multiple of $\mathbf{f}$. The opportunity for roguery arises when we consider $\mathbf{f} = \alpha \mathbf{S}_X + \beta \mathbf{S}_Y + \gamma \mathbf{1}$. Since $\mathbf{v} \cdot (\alpha \mathbf{S}_X + \beta \mathbf{S}_Y + \gamma \mathbf{1}) = \alpha S_X + \beta S_Y + \gamma$, each player can unilaterally enforce a linear relation between the players' long term average payoffs by setting the determinant to zero. The strategies that enforce this linear relationship were christened *zero determinant (ZD)* strategies.

Press and Dyson focused on two types of ZD strategies. The first allows X to unilaterally set her opponent's score by setting $\alpha = 0$ and thus forcing $\beta S_Y + \gamma = 0$. Solving the equations

$$-1 + p_1 = \beta \mathrm{R} + \gamma$$
$$-1 + p_2 = \beta \mathrm{T} + \gamma$$
$$p_3 = \beta \mathrm{S} + \gamma$$
$$p_4 = \beta \mathrm{P} + \gamma$$

to eliminate the parameters β and γ and expressing p_2 and p_3 as functions of p_1 and p_4, they arrive at the strategy

$$p_2 = \frac{p_1(T - P) - (1 + p_4)(T - R)}{R - P}$$
$$p_3 = \frac{(1 - p_1)(P - S) + p_4(R - S)}{R - P}$$

which sets Y's score to a weighted average of P and R

$$S_Y = \frac{(1 - p_1)P + p_4 R}{(1 - p_1) + p_4}.$$

This equation has feasible solutions when p_1 is close to but not equal to 1, and p_4 is close to but not equal 0.

Using a similar calculation Press and Dyson show that X cannot set her own score with $\tilde{\mathbf{p}} = \alpha \mathbf{S}_X + \gamma \mathbf{1}$, as

$$p_2 = \frac{(1 + p_4)(R - S) - p_1(P - S)}{R - P} \geq 1$$

except for the limiting case $\mathbf{p} = (1, 1, 0, 0)$.

The second type of ZD strategies described by Press and Dyson is even more devilish. It allows X to demand and get an extortionate share of the payoff surplus over the mutual defection value P by setting

$$\tilde{\mathbf{p}} = \phi[(\mathbf{S}_X - P\mathbf{1}) - \chi(\mathbf{S}_Y - P\mathbf{1})]$$

which leads to

$$\begin{aligned} p_1 &= 1 - \phi(\chi - 1)\frac{R - P}{P - S} \\ p_2 &= 1 - \phi\left(1 + \chi\frac{T - P}{P - S}\right) \\ p_3 &= \phi(\chi + \frac{T - P}{P - S}) \\ p_4 &= 0 \end{aligned}$$

with feasible strategies existing for any χ and

$$0 < \phi \leq \frac{P - S}{(P - S) + \chi(T - P)}.$$

Computing the long-run average scores for X and Y with using Axelrod's values $(T, R, P, S) = (5, 3, 1, 0)$ Press and Dyson show

$$S_X = \frac{2 + 13\chi}{2 + 3\chi} > 3 = P, \text{ and } S_Y = \frac{12 + 3\chi}{2 + 3\chi} < 3 = P$$

for $\chi > 1$, while the limiting fair case $\chi = 1$ and $\phi = 1/5$ reduces to TIT FOR TAT with $\mathbf{p}^{\text{TFT}} = (1, 0, 1, 0)$.

When playing against an extortioner X, Y has two choices, the first is to try to maximize his own score, but that would feed into X's plot and increase X's score even more. The second is to refuse to be extorted by playing AllD either out of spite or in hopes that X recognizes that extortion has failed and switches to a more equitable play. The key to understanding a match between two extortioners lies in their reaction to mutual defection given by $p_4 = q_4 = 0$. Once they reach the state DD, they will stay there indefinitely thus receiving a long term average of P. This should raise questions about evolutionary stability of extortionate ZD strategies. In the two years since Press and Dyson published their findings, a number of papers came out discussing how well can ZD strategies fare in ecological and evolutionary settings. Before moving on to those results, however, I should note two more theorems established by Press and Dyson.

The first shows that restricting our attention to memory-one strategies is not as limiting as it may seem. They prove that if X plays a memory-one strategy and Y plays a longer, but still finite, memory strategy, then there exists a memory-one strategy for Y that would lead to equivalent payoff to the longer memory strategy. In evolutionary setting, it is conceivable that longer memory may still be useful. In their 2012 paper Adami and Hintze [2] note that longer memory may provide evolutionary advantage to variants of ZD players who attempt

to recognize each other by analyzing longer pattern of play and switch to mutual cooperation in order to increase their evolutionary fitness.

The second ancillary result dispels doubts that may arise about the steady state payoff $S_i = \mathbf{v} \cdot \mathbf{S_i}$, $i \in \{X, Y\}$. It may take hundreds of moves for the linear relationship

$$\alpha S_X + \beta S_Y + \gamma = 0$$

to be established, and one may wonder if there is any way the non-ZD player can prevent that relationship by changing its play inside the Markov equilibration time scale. Press and Dyson answer that question negatively, thus showing that in a one-on-one play the non-ZD player cannot benefit from having longer memory or by keeping the game away from the Markov stationary distribution. Thus they set the course for the future research to be focused on memory-one strategies, which significantly simplifies the space of strategies and the search for its structure.

What can ZDs do?

Since ZD strategies may take hundreds of turns for the average payoffs to settle near their expected values, they may not be particularly useful in modeling situations where interactions are not frequent or the expected number of interactions is low, like human behavioral experiments or duopoly output where companies have delayed feedback about their opponents. It is easier to imagine these behaviors to arise in evolutionary or ecologic systems where numbers of interactions over thousands of generations may be much higher and the laws of large numbers take effect. One needs not assume that players in evolutionary systems are aware of ZD strategies and are capable of consciously calculating the probabilities (p_1, p_2, p_3, p_4) if ZD strategy could evolve as a result of natural selection. The possibility of evolutionary dynamics giving birth to ZD strategies and their stability once they are introduced to a population is the subject of a number of research papers published shortly after Press and Dyson made their discovery. The remainder of this paper presents some of these results.

Within a few month of Press and Dyson's publication, Stewart and Plotkin [13] performed a simple experiment. They recreated Axelrod's first round-robin tournament with several additional players: GENEROUS TIT FOR TAT (GTFT), TIT FOR TWO TATS (TF2T) (recall that TF2T would have won Axelrod's first tournament had it been submitted), the compliant *zero determinant generous tit for tat 2 (ZDGTFT-2)* which forced the relationship $S_X - R = 2(S_Y - R)$ so that $S_X \leq S_Y$, and *Extort-2* with $\mathbf{p}^{\mathrm{E2}} = (8/9, 1/2, 1/3, 0)$. As one may expect, Extort-2 won the highest number of matches save for AllD. Those victories, however, were mostly pyrrhic. Extort-2's average score was second to last (again bested, or rather worsted, only by AllD). More surprisingly ZDGTFT-2 received the highest average score followed by Axelrod's usual suspects GTFT, TFT, and TF2T.

Stewart and Plotkin's simulations suggest that extortioners do not do well when faced with a population consisting of a variety of strategies. This hypothesis was further explored by Christoph Adami and Arend Hintze in a paper that first appeared in August 2012 and has since undergone several revisions [2]. They analyzed evolutionary performance of the two types of ZD strategies discussed by Press and Dyson, the dictatorial equalizer ZD^D that unilaterally sets the opponent's score, and the extortioner ZD^E. Except for the limiting

case with $p_1 = 1$, the dictatorial ZD^D strategies do not fare well against their own copies, as they force on their clones the same score $W(ZD^D, ZD^D) \leq R$ that they force on all other strategies, and thus are evolutionarily dominated by TFT, Pavlov, and other strategies that receive a score of R against their own kind. Extortioners suffer from even poorer performance against other extortioners, since $p_4 = 0$ inexorably drives them to mutual defection and the long term average of P. In two-strategy competitions some strategies do give ZD^D the upper hand. In particular the general cooperator GC which Iliopoulos, Adami, and Hintze showed to be a fixed point at low mutation rates [8] is dominated by ZD with $W(Z, GC) = 2.125$ and $W(GC, GC) = 2.11$. However this advantage does not contradict their earlier findings. Agent based simulations with mutation rate favoring GC and seeded with the dictatorial ZD ($p_1 = .99$, and $p_4 = .01$) showed that the ZD^D strategy evolves into GC [2], and is thus evolutionarily and mutationally unstable. In closing, Adami and Hintze concede that extortionate ZDs may be evolutionarily fit if they learn to recognize each other either through a tagging mechanism or by having access to longer play history (recall that ZDs are all memory-one strategies), and adapt to cooperate with other ZDs and extort from non-ZD players. But they note that other strategies could take advantage of that recognition by developing fake tags or patterns of play that would make them appear ZD and thus provoke cooperation. Nature provides countless vibrant examples of such adaptation including the familiar Syrphidae flies that mimic the coloring of bees or wasps and thus protect themselves against predators.

In a 2013 paper Christian Hilbe, Martin A. Nowak, and Karl Sigmund analyzed a special case of prisoner's dilemma called the donation game with the payoff structure given by

$$\begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix}$$

and the pool of strategies consisting of the zero determinant extortioner with extortion factor χ (ZDE_χ), TFT , $WSLS$, $AllC$, and $AllD$. They note that in pairwise comparisons E_χ and $AllD$ are neutral, TFT weakly dominates E_χ by performing better against other TFT than E_χ plays against other extortioners, and that $AllC$ and extortioners can invade each other and stably coexist in proportion $c(\chi - 1) : (b + c)$, and finally $WSLS$ dominates E_χ. TFT can always invade mixed equilibria of extortioners and unconditional cooperators or defectors, but can in turn be invaded by other nice strategies.

To study equilibrium distributions of these strategies in finite populations, Hilbe et al. modeled natural selection as an imitation process. At each step two randomly chosen players X and Y compare their average payoffs S_X and S_Y, and Y switches to X's strategy with some probability that is a function of the difference $S_X - S_Y$. Although in this model extortioners were never the most abundant outcome, they played an interesting role of catalyzing cooperation. Populations made up solely of $WSLS$ and $AllD$ were stuck in stable equilibria containing predominantly $AllD$, however when E_χ or TFT were added to the mix, equilibrium distributions favored $WSLS$ in all but the smallest populations. The results were similar with the limiting case of rare mutations (with equilibria computed analytically using the method of Fudenberg and Imhof [7]) and more frequent mutations (equilibria computed by agent-based simulations). Except in small populations, rare mutations led to $WSLS$ fully overtaking the population and driving $AllD$, E_χ, $AllC$, and TFT extinct. When mutations were more frequent, the populations stabilized at mixtures heavily favoring $WSLS$, but not

driving others to complete extinction. In these scenarios extortionate ZDs (including the fair TFT) were shown to be catalysts of cooperation. They interestingly complement and deepen the previous results discussed here, including Axelrod's chronology of emergence of cooperation [3, p.55] and long-term dominance of WSLS discovered by Nowak and Sigmund [11]. It is also a partial rebuttal to Adami and Hintze's assertion that 'winning isn't everything' [2], as Hilbe et al. show that evolutionarily unstable strategies can nonetheless play a vital role in evolution of cooperation.

Hilbe et al. present another scenario in which extortioners prosper. When the IPD is played between two populations (hosts and their symbionts), or two types within the same populations (buys and sellers, males and females, constituents and representatives), extortion strategies can evolve even in large populations because their low scores when playing their own type does not hinder their fitness. If the first group plays E_χ and the second group attempts to maximize their own score by playing the best response, the second group will adopt $AllC$. However in this situation the first group is likely to evolve an even more profitable $AllD$ which would force the second group to also switch to $AllD$. An interesting special case was revealed by simulations in which the two populations evolved at different rates of mutations. In a simulation seeded with non-ZD strategies the slowly evolving host populations arrived to a $\delta = 0.1$ neighborhood of E_χ (Euclidean distance on the 4-cube (p_1, p_2, p_3, p_4)), and were able to extract a surplus more than 10-fold larger than what was achieved by the rapidly evolving symbionts. The hosts' payoff after 2000 generations was consistently above the mutual cooperation value $R = b - c = 2$, while the symbionts received scores close to $P = 0$.

The discovery of ZD strategies by Press and Dyson motivated a number of papers that focused on extortionate behavior, however it could be that the more successful ZDs are of an entirely different type – the generous compliers. Whereas extortioners offer their opponents a choice between receiving a low long-term score of P by maintaining mutual defection or letting the extortioner reap an unfair share of the surplus, the generous ZD proposes that both players receive the long-term score of R and suffers a greater score loss than the opponent if they deviate from mutual cooperation. In their commentary to Press and Dyson, Stewart and Plotkin showed that a generous strategy they called ZDGTFT-2 won an Axelrod-style tournament that contained TFT, GTFT, TF2T, WSLS and a number of other strategies [13]. Using the criterion of *evolutionary robustness*, a modification of ESS to finite populations, they explored the subset of generous ZD strategies (of which ZDGTFT is an example) and found that generous ZDs perform very well in simulations, and in some conditions even outperform WSLS [14]. The chronology of research following Press and Dyson's introduction of ZD strategies is reminiscent of Axelrod's tournaments. Following the success of TFT the players were looking to gain an upper hand by submitting variants of TFT that attempted to take advantage of opponents by not always being nice. These versions almost invariably did worse than TIT FOR TAT, but also worse than more generous variants TF2T and GTFT. Similarly the initial interest in ZDs focused on extortionate and dictatorial equalizer strategies designed to gain at cost to opponent, when in reality ZD strategies that practiced generosity, the opposite of extortion, turned out to give their practitioners better evolutionary outcomes. I find that these results offer an optimistic view of nature, and a somewhat pessimistic view of humanity as represented by the community of evolutionary game theorists whose first instinct, it seems, is to not be very nice.

3.1 Notation

π_i payoff to player i in a one-shot game

$W(I, J)$ expected long-term payoff for strategy I playing against strategy J

w the probability of repeat interaction in an intreated game

$AllD = (0, 0, 0, 0)$ the strategy of always defecting

$AlLC = (1, 1, 1, 1)$ the strategy of always cooperating

$TFT = (1, 0, 1, 0)$ the strategy of cooperating on first move and repeating opponents previous action on each subsequent move

$WSLS$ or $PAVLOV = (1, 0, 0, 1)$

Bibliography

[1]

[2] Christoph Adami and Arend Hintze. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nature Communications*, pages –, 2014.

[3] R. Axelrod. *The evolution of cooperation*. Political Science / Science. Basic Books, 1984.

[4] Robert Axelrod. The emergence of cooperation among egoists. *The American Political Science Review*, 1981.

[5] Robert Boyd and Jeffrey P. Lorberbaum. No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature*, 327(6117):58–59, May 1987.

[6] R. Dawkins. *The Selfish Gene: 30th Anniversary Edition*. ISSR library. OUP Oxford, 2006.

[7] Drew Fudenberg and Lorens A. Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131(1):251–262, November 2006.

[8] Dimitris Iliopoulos, Arend Hintze, and Christoph Adami. Critical dynamics in the evolution of stochastic strategies for the iterated prisoner's dilemma. *PLoS Comput Biol*, 6(10):e1000948, 10 2010.

[9] J. Maynard Smith and G. R. Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

[10] J.F. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[11] Martin Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432):56–58, 1993.

[12] R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55, 1975.

[13] Alexander J. Stewart and Joshua B. Plotkin. Extortion and cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 109(26):10134–10135, 2012.

[14] Alexander J. Stewart and Joshua B. Plotkin. From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 2013.

[15] Gerald S. Wilkinson. Reciprocal food sharing in the vampire bat. *Nature*, (5955):181?184, 1984.