PROBABILISTIC LEARNING AUTOMATA AND THE PRISONER'S DILEMMA

Ath. Kehagias

I. Introduction

The Prisoner's Dilemma (henceforth PD) has been discussed extensively as a model of the conflict between competition and cooperation, or between individual and collective rationality. PD can be viewed either as a mathematical paradox, or as a useful (albeit simplified) paradigm that belongs to the realm of applied sciences such as psychology, economics, political science, biology, ecology etc. A good technical presentation of PD appears in (Rapoport, 1966). A bibliography and discussion of the technical aspects can be found in (Luce & Raiffa, 1985); an extensive treatment of variations of PD can be found in (Axelrod, 1984). For an interesting popularized discussion see (Hofstadter, 1985).

Probabilistic Learning Automata (henceforth PLA) are an important model of Artificial Intelligence. They offer a simple way to describe the learning process of several interacting agents. In this sense they are a very suitable model for the evolution of strategies for playing PD. For a very extensive discussion of the theory of PLA, as well as for a presentation of their applications see (Narendra & Thathatchar, 1989).

In this paper we discuss an iterated version of the PD , played by Probabilistic Learning Automata (henceforth PLA). This particular version of PD has been studied by computer simulation; we summarize our findings and discuss possible implications. The main conclusion is that cooperation is a more viable and persistent alternative than competition. This is an intuitively satisfying result and may help in resolving the apparent paradox of PD. Our results must be viewed wth caution; strictly speaking they only pertain to the particular version of PD that we discuss here. However, even this limited analysis may offer useful insights to more general versions of PD and to the general competition - cooperation problem. A detailed presentation of the computer simulations and the associated mathematical analysis is outside the scope of this journal and will appear elsewhere.

II. Prisoner's Dilemma

Consider the following situation. Two persons, call them A and B, are accused of having committed a crime. The prosecutor makes A the following proposition. "There is circumstantial evidence against both you and your accomplice. If both of you plead innocent you will be convicted anyway, and each will receive a sentence of two years imprisonment. However, if you give evidence against your accomplice, we will have a better case against him and he will be convicted to a harder sentence: five years imprisonment. In exchange for your help, we will let you free for turning state's evidence." The prosecutor makes B exactly the same proposition; he also tells both A and B that the proposition was made to the other person. Finally, he tells them that if they both confess each will receive a sentence of four years. A and B cannot communicate with each other; each must decide independently whether to denounce his partner, or cooperate with him. This is the Prisoner's Dilemma. It will shortly become clear why it is called a dilemma.

We can view PD as a game to be played between A and B (with the prosecutor acting as referee). Hence we will use the following terminology. A and B will be called players or agents. Each player has a choice between two actions: C (cooperating with the other player) and D (defecting or denouncing the other player). The outcome of the game (the resulting sentences) will be called the cost of the actions. Finally, we will assume that each player has only one criterion for playing the game, namely to minimize his cost, without being concerned with questions of trust, ethics, friendship etc. We will discuss this assumption in greater detail later.

A preliminary analysis of the PD game might go as follows. A is tempted to defect (denounce B) and go away free, but he realizes that B is under exactly the same temptation and might denounce him in turn. In a case of a double defection, both A and B will be convicted to four years imprisonment, which is clearly undesirable. If A and B cooperate, they get the lighter sentence of two years each; but since they cannot communicate, each of them must choose his action independently. The analysis is facilitated by introducing the following tables. Table 1 summarizes the cost of each choice to both players.

B Actions A Actions	C	D
C	2 , 2	5 , 0
D	0 , 5	4 , 4

Table 1

B Actions A Actions	C	D
C	2	5
D	0	4

Table 2

B Actions A Actions	C	D
C	2	0
D	5	4

Table 3

A's action determines a row and B's action determines a column of Table 1; then the entry at the chosen column and row determines the cost to A (first number) and to B (second number). Table 2 summarizes the cost to A only and, similarly, Table 3 summarizes the cost to B only. Note that each player's cost depends on both players' actions.

We have assumed that each player's sole objective is to minimize his cost, hence considerations of friendship, trust, morality etc. are assumed to be irrelevant. However, if we wished, we could introduce such factors to the game by assigning them a cost. For example, if defection is a "bad thing", we can increase the cost of all D actions by a certain amount, say 2. Of course this presumes that moral choices can be assigned a numerical value; this is, indeed, a bold assumption. At any rate, from the players' point of view, the only thing that matters is minimization of cost and hence Table 1 is, at least in principle, a complete description of the PD game.

Let us now return to the analysis of the game. When A chooses his action, he looks at Table 2 to compute the respective cost. He can reason as follows. Suppose that B chooses action C (i.e. the first column of Table 2). Then A can choose action C, at a cost of 2, or action D at a cost of 0. Since $0 < 2$, it is preferrable to choose action D. Similarly, suppose that B chooses action D (i.e. the first column of Table 2). Then A can choose action C, at a cost of 5, or choose action D at a cost of 4. Since $4 < 5$, it is again preferrable to choose action D. Thus, no matter what action B chooses , it is best for A to choose action D. But the game is perfectly symmetric. B will choose exactly the same action D, by exactly the same reasoning; in which case both A and B incur a cost of 4 (four year prison sentences). Here lies the paradox: if they had both chosen C, they would only incur a cost of 2. So, by seemingly faultless reasoning they choose an action which is clearly suboptimal for both of them.

Of course A might take his reasoning one step further, perceive the cost of a mutual defection and conclude that it is better to trust B and choose action C. But this conclusion is only valid assuming that B will also choose C; in which case A would be better off by choosing D anyway. Hence, it seems there is no way out of this vicious circle which enforces noncooperative D actions. This is the Prisoner's Dilemma. It is a dilemma in that, while cooperation is a clearly preferrable alternative, it cannot be justified; no for each player competitive behavior appears to yield a lower cost, matter what the other player does.

III. Further Examples

At first sight PD appears to be a frivolous and contrived problem, but a little reflection will show that it is a simplified version of many real world situations.

For example, think of an arms race between two rival countries. Each country has the options of arming or not arming. Assume that both countries can arm at the same rate. If both countries arm neither gains an advantage over its rival; in addition they are both burdened by an increased military budget. Clearly it is preferrable that neither country arms. On the other hand, if one country arms and its rival doesn't, then the first country gains an advantage. Clearly, arming corresponds to the D move of the previous section and not arming corresponds to the C move. The best situation for a country is that it arms and its rival doesn't; barring this, its is best that neither country arms; the worst situation occurs when both countries arm. However, just as in the prisoners' case it is clear that no matter what one country does it is best for the other country to arm. Or is it? This is a simple but not inaccurate model for the Cold War rivalry between the US and USSR. Virtually the same analysis can be applied to intercommunal conflicts such as the one between Protestants and Catholics in Northern Ireland, Jews and Palestinians in Israel, Moslims and Serbs in Bosnia and so on. In some of these cases the deadlock has been resolved and a cooperative solution found; in other cases the "players" are stuck with noncooperative strategies.

In a different context, consider an enviromental problem. The water resources of a city are running low. There is a steady but low inflow of water, which is not sufficient for the needs of all the population. To simplify matters, assume that the city has a population of only two people. Each one of them has the choice of conserving water (which will be inconvenient but not unbearable) or consuming at a high rate. If both consume at a high rate, pretty soon the city reservoirs will be empty and the population thirsty. If both conserve, the water resources will increase and the crisis averted. And if only one consumes, the other can use all the water he wants and the water resources will remain constant. Once again each "player" has an incentive to consume, no matter what the other does, but this leads to a catastrophic outcome. If the population were not two people but one million, this analysis would apply fairly well to the plight of a certain Greek city. Of course the analysis is not limited to water; clean air, oil and a number of other resources (even parking space at city center) could be used instead.

Similar examples can be found in many other areas: economics (competing firms), biology (competing organisms) and so on.

IV. Generalizations and Discussion

Admittedly all such examples are much more complex than the PD. But the essence of the dilemma is captured in a simple table of action costs, such as Table 1. In fact, a slightly more general table, such as Table 4, is appropriate.

B Actions A Actions	C	D
C	R , R	S , T
D	T , S	P , P

Table 4

Here R, S, T, P are the costs of each pair of actions. To have a PD situation it is sufficient that the following inequalities hold.

(1a) $T > R$: If B cooperates, it is better for A to defect (and vice versa).
(1b) $R > P$: If both A and B cooperate, they do better then if both defect.
(1c) $P > S$: If B defects, it is better for A to defect (and vice versa)

When (1a-c) hold we have a situation where both players are tempted to defect; but if they both defect they are worse off then if they both cooperated. For either player, the best situation is when he defects and his opponent cooperates. Hence there is temptation for individual defection, but mutual defection is costly for both players. The crux of the matter is that "selfish", competitive reasoning leads to a result that is bad from both the individual and collective point of view, while "altruistic", cooperative reasoning leads to a result that is good from both the individual and collective perspective. The simplified PD model captures the essential elements of this situation and may yield useful insights into more complex, real world problems. Undoubtedly, this an important reason for the great interest PD has aroused.

Another reason for the interest in PD must be a sense of frustration. While cooperation is obviously better than defection, the reasoning for defection appears to be faultless. This is morally distasteful to many people; perhaps more importantly, experimental observation of actual human PD players shows that very often they will play cooperatively. Therefore, there is a strong incentive to find a rational way out of the dilemma.

Many authors have attempted to resolve PD by introducing in the decision process motives such as ethics, trust, friendship etc. This can be done in two ways. The simplest way has already been mentioned: it is to assign to such "higher motives" a numerical cost. For example, if defection is morally reprehensible, add, say, ten units to the cost of a defecting player and build a new table, which incorporates the additional costs. But this only postpones the dilemma. In some instances of PD, the additional costs might produce a table that does not satisfy inequalities (1a-c), for example when defection incurs a high moral cost. But it may be that unreciprocated cooperation also has a high cost, or succesful defection has a high profit. For instance, returning to the prisoners' example, imagine that the penalty for the lone cooperator is not five years of prison, but the death sentence; in this case it takes a very trusting player to play C. Hence, despite the moral cost, situations will arise where (1a-c) still hold and these will introduce the same paradox as in the original PD.

A more serious objection (and a possible way out of the dilemma) is that there are situations in real life which cannot be assigned a numerical cost. What is the cost of death? What is the cost of ostracism when one defects? What units are these costs measured in? Are they the the same units for death and ostracism? These are serious objections and, in our opinion, have not been answered in a satisfactory manner yet.

Hence one can escape the paradox of PD by rejecting its applicability to real life: the paradox never arises, because people do not assign numerical costs to their actions and do not determine their actions according to such costs. But we think this is not a very good argument, or at least it does not make the study of PD uninteresting. It is probably true that very few people will follow an exact PD analysis in their everyday decision making, but we believe that game tables such as Table 1 are acceptable approximations. While they cannot give exact predictions of the way people act, they may help in qualitatively understanding certain motives of human behavior.

At any rate, this is a problem for the psychologist and behavioral scientist. Let us now present a different attempt to escape the PD paradox, which can be studied mathematically.

Many people have accepted the assignment of numerical costs to PD-like games, but have observed that such games are rarely played only once. What occurs more frequently is the so called iterated PD. The PD game will not be played once, but many times in succession, and each player will remember how his opponent played in the past. This can promote cooperation, both in a negative and a positive way. For example, we have established that the "individual-rational" way to play PD is for both players to defect. This happens in the first round of PD and consequently both players are punished. This may convince them to be a little more cooperative in the next round. Conversely, suppose both players cooperate (even by accident) once; consequently they are rewarded and this may make them more cooperative in the future. The rest of this paper is concerned with formalizing this argument and exploring its consequences.

A related idea that becomes more plausible when one considers iterated PD is the use of mixed strategies. One could decide to play C, say, 50% of the time and D 50% of the time. This decision when to play C and when D could be taken deterministically, e.g. by always cooperating once, then defecting, then cooperating again and so on, or probabilistically: e.g. by flipping a coin at every round and cooperating if it comes up heads, defecting otherwise. In general one could cooperate with probability p and defect with probability q, where p+q=1. The decision when to cooperate and when to defect can be taken using a computerized coin that can come up heads with any desired probability p and tails with any desired probability q.

Before concluding this section, let us mention a final possibility for generalizing the PD. In everything we have discussed so far, we have assumed that only two players are involved. But one can also consider the case of three-, four- or, in general, N- player PD. In such a case, the cost to each player depends on the action of several other players. One can set up a cost function (rather than cost table) which depends on the actions of several players, and which yields large costs when many players defect, and small costs when many players cooperate; in addition the function is such that the optimal situation for a single player (smallest cost) occurs when everybody else cooperates and he defects. Hence each individual player has an incentive to defect, but when many players defect a large cost is inflicted to everybody. Obviously, such an N-player PD is a more realistic model for the environmental problems we discussed earlier. The introduction of N players creates many additional complications in the analysis of the game; indeed the study of N- player games is considerably harder than that of two- player games. At any rate, we will not discuss this possibility in this paper; it is just mentioned for the sake of completeness.

V. Probabilistic Learning Automata

In this section we introduce PLA's, which we have used in our study of the iterated PD. We will present a very brief summary of the main ideas; for a complete discussion see (Narendra & Thathatchar, 1989). The basic idea is to develop a model of the learning process for a group of individual simple agents (automata) that repeatedly interact with each other.

Consider an automaton (in other words a very simple entity) that interacts with the surrounding environment in the following manner. At time instants t = 1, 2, ... etc. the automaton chooses one out of two possible actions; as a consequence of the chosen action, it receives from the environment a response, which can be either a reward or a punishment. The connection to the PD is obvious: take the automaton to be a PD player, the possible actions to be C and D, and the cost of each action to be the corresponding prison sentence; a short (or zero) sentence is a reward, a long sentence is punishment.

Generally, the automaton should choose its actions in a way that increases reward and minimizes punishment; in other words it should try to minimize the cost of playing the game. This cost minimization has two aspects. On the one hand the automaton tries to maximize its immediate rewards; on the other hand it tries to learn the behavior of the environment (and of other automata), and use this knowledge for selecting its future actions. If the same action were chosen always, the response and the cost would also remain the same. Rather, one wants to use the previous actions and responses, in combination with some learning scheme, to decrease cost as time progresses. We will now present such a scheme.

First, our automata will use mixed strategies, in other words they will choose their actions probabilistically. For instance, at every time step the automaton flips a coin. If the coin comes up heads, the automaton chooses C; otherwise it chooses D. Obviously, the probability of C is p =1/2 and the probability of D is q = 1/2. Or the automaton can choose actions by tossing a die. If a six comes up, C is chosen; otherwise D. In this case p = 1/6 and q =5/6. Using a "computer coin" (more precisely, a computerized random number generator), the automaton can choose actions with any desirable probabilities p and q.

If the action probabilities were constant and independent of the enviroment's response, then no learning would occur. The one- round and iterated PD games would have exactly the same outcome. Instead our automata will use time varying action probabilities $p(t)$ and $q(t)$. The probability of choosing C at time t is $p(t)$ (and similarly for $q(t)$). Further, $p(t)$ and $q(t)$ change according to the environment's responses. For instance, if at time t the automaton chooses C and the environment rewards it, then the probability of choosing C is increased for subsequent time steps. A simple scheme that implements this idea is the following.

(2a) $p(t+1) = p(t) + (1-p(t))\cdot á$ if at t action = C, response= reward,

(2b) $p(t+1) = (1-á)\cdot p(t)$ if at t action = C, response = punish.

Here a is a number representing learning rate. When a is large, p (t) can change a lot even in one time step. Conversely, when a is small, at every time step $p(t)$ changes only a little. Hence the first equation tells us that when C is chosen and rewarded, then the probability of choosing C at the next time t +1 is increased by a positive quantity; when the action is punished the probability is multiplied by a number less than 1, hence it decreases. Recall that there are only two possible actions, with probabilities p and q respectively. Hence for every t we must have $q(t) = 1 - p(t)$. So equations (2a-b) imply two complementary equations for updating $q(t+1)$.

In case action D is chosen, similar equations are used for updating the action probabilities:

(3a) $q(t+1) = q(t) + (1-q(t))\cdot á$ if at t action = D, response= reward,

(3b) $q(t+1) = (1-á)\cdot q(t)$ if at t action = D, response = punish.

Again, (3a-b) imply two complementary equations for updating $p(t+1)$. Equations (2a-b, 3a-b) constitute a simple scheme that will learn connections between actions and responses. For example, consider the case where C is always rewarded and D always punished. Suppose that our automaton starts with equal probability of choosing the first or second action ($p(1) = q(1) = 1/2$) and chooses 100 consecutive actions; further suppose that 50 of these are D actions. Since we assumed that D is always punished, after 100 time steps we will have $q(100) = (1-a)^{50} \cdot 1/2$. For a =0.9, this probability would be about 10^{51}, which means that by time t = 100, the D action probability is very close to zero and the automaton will practically never choose D any longer. So the PLA has learned that C is "good" and D is "bad" .

VI. A PLA, Iterated PD Game

Now we have all the pieces necessary to design an experiment of iterated PD. We will use not one, but two PLA's, one as player A and the other as player B. At time t , automaton A chooses C (cooperation) with probability $pA(t)$ and D (defection) with probability $qA(t)$; similarly automaton B chooses C with probability $pB(t)$ and D with probability $qB(t)$. Then the automata communicate their chosen actions to an impartial referee (the "environment"), who responds by meting out reward and punishment according to the PD cost table. Finally, the automata update their action probabilities according to the response they received.

For example, suppose A chooses C and B chooses D. Then A receives a cost of S and B receives a cost of T. Given B's action, A notices that, it would have better had it chosen D. In particular A notices an opportunity loss of S-P. Hence it perceives the environment response as punishment and it computes its loss as S-P (in other words the cost it actually incurred minus what it would have incurred had it chosen D) and decreasess its cooperation probability by equation 3b, using a learning rate a=(S-P). Similarly, B perceives a gain of R-T and increases its defection probability by a learning rate of a=(R-T). Similar reasoning applies in the case that the automata choose actions CC, DC or DD. The only modification from the standard learning method of the previous section is that the learning rate a is not fixed but variable, depending on the opportunity cost of the differences S-P, P-R, R-T.

In this way we have devised a scheme such that each automaton continuously updates its action probabilities, depending on the actions of itself and its opponent's. It is a fairly simple scheme, but it can be used to explore the evolution of cooperation and competition under various conditions, which are determined by the choice of parameters P, R, S, T. (Several other parameters are available for experimentation, but will not be considered here, because of space limitations).

VII. Computer Experiments

Hence our experiment plan is the following. For each experiment we choose specific T, R, S, P parameters and we run a computer simulation of equations (2a,2b,3a,3b3) for a a large number of time steps t =1, 2, ... We take enough time steps to ensure that the learning process is completed and observe the final values of the cooperation and defection probabilities. There are four such probabilities: *pA* (cooperation probability for automaton A), *qA* (defection probability for automaton A), *pB* (cooperation probability for automaton B), *qB* (defection probability for automaton B). The *pA*, *qA*, *pB*, *qB* values at the end of experiment describe the level of cooperation achieved. For instance we could have *pA*=1, *pB*=1; in this case both A and B would choose the cooperative move 100% of the time; or we could have *pA*=*pB*=0, in which case A and B will choose the defection move 100% of the time; or *pA*=1, *pB*=0.5, in which case A will always choose C, but B will choose C 50% of the time, D 50% of the time and so on.

We expect that the final level of cooperation, as expressed by pA, pB, qA, qB, will depend on the values of the costs T, R, P, S. More precisely, since the learning process described in the previous section depends not on T, R, P, S themselves but on the differences S-P, P-R, R-T, we also expect that in each experiment the final *pA*, *pB*, *qA*, *qB* values are determined by S-P, P-R, R-T. By adjusting these values we can promote cooperation or competition. For example, if T is much smaller than R, then on every C move of B, A loses much more by cooperating than by defecting. In this case defection is promoted. If P is large compared to R, mutual defection has a much larger cost than mutual cooperation, hence cooperation is promoted, and so on. There is considerable latitude in the choice of P, R, S, T, but to have a PD game we must alwaysrespect inequalities (1a-c).

We have run several experiments along these lines, trying various combinations of P, R, S, T values. The results are presented in Table 5.

We note that in some cases we obtain a pure cooperation strategy and in some others a mixed strategy, but never a pure defection strategy. In general, large P and R values promote cooperation and large S and T values promote defection, as expected. In short, Table 11 tells us that cooperation is always a viable strategy, and when the T, R, P, S are closely spaced, a pure cooperation strategy will emerge.

$p_A(1)$	$p_B(1)$	$\acute{a}$	R	T	S	P	$p_A(700)$	$p_B(700)$
0.50	0.75	0.010	1.0	0.0	5.0	4.0	1.000	1.000
0.50	0.75	0.010	1.0	0.0	5.0	2.0	0.378	0.375
0.50	0.75	0.010	1.0	0.0	5.0	4.5	1.000	1.000
0.50	0.75	0.010	1.0	0.0	2.0	1.5	0.384	0.434
0.50	0.75	0.010	2.0	0.0	5.0	4.0	0.617	0.631
0.50	0.75	0.010	3.0	0.0	5.0	4.0	0.338	0.396
0.50	0.75	0.010	3.5	0.0	5.0	4.0	0.263	0.265
0.50	0.75	0.010	1.0	0.0	9.9	4.0	0.462	0.499
0.50	0.75	0.010	0.1	0.0	4.1	4.0	1.000	1.000
0.50	0.75	0.010	0.1	0.0	5.0	4.0	1.000	1.000
0.50	0.75	0.010	0.1	0.0	9.9	4.0	0.596	0.597
0.50	0.75	0.010	3.9	0.0	4.1	4.0	0.116	0.119
0.50	0.75	0.010	3.9	0.0	5.0	4.0	0.108	0.108
0.50	0.75	0.010	3.9	0.0	9.9	4.0	0.088	0.088

Table 5

VII. Conclusion

The results of our computer experiments are interesting, but some caution is necessary: these experimental results do not constitute a complete analysis of our model. Such an analysis would require inordinate amounts of computer time to fully explore all possible combinations of parameters. Another method would be a mathematical analysis of convergence properties of our model. Such an analysis requires rather sophisticated mathematical methods and belongs to a more specialized journal. For an exposition of such methods see (Narendra & Thathatchar, 1989).

Even a complete mathematical analysis will only give information about this particular model of playing iterated PD. Our model is a linear reward - penalty probabilistic learning automaton. Many other types of learning laws could have been used, such as reward-inaction, inaction-penalty, linear or nonlinear and so on. For a full discussion of learning laws for PLA's see (Narendra & Thathatchar, 1989). Further, learning automata are only one of many possible models for playing iterated PD. Finally even a complete analysis of every possible strategy for playing iterated PD (which appears to be a formidable task) would still pertain to PD; it must be stressed that PD is a vastly simplified model of realistic cooperation - competition situations.

Keeping all these qualifications in mind, we still have evidence for the following rather optimistic conclusion: in a PLA game of iterated PD cooperation will emerge under very broad conditions and, excluding the case of very low defection cost, we will in fact have a pure cooperation strategy. Also, the behavior of our model is intuitively appealing, since its dependence on the cost parameters is generally the one we would expect. For instance, we observe that increased T hinders, or at least delays cooperation, while increased R and P promote it.

These conclusions contradict a number of more pessimistic analyses of PD existing in the literature. We may have here at least a partial resolution of the PD paradox. It seems that the crucial elements are iterated playing combined with learning.

A more detailed study, involving extensive computer experimentation and mathematical analysis of our model will undoubtedly clarify the issue further. Other extensions of the model could include the application of the PLA/PD framework to more realistic problems of competition and cooperation, in particular N player problems. For example we could have a macroeconomic model of investment and consumption: several consumers share a resource (capital) that can be either consumed or invested; one would look for an optimal policy that maximizeseach consumer's utility. Would such a policy be competitive (immediately consume as much as possible) or cooperative? Finally, a very ambitious research goal would be the empirical test of our PLA/PD model, involving the observation of humans playing iterated PD. Records of actions and costs can be used to statistically estimate the T, R, P, S parameters, so that a computer experiment will replicate as closely as possible the observed human behavior.

References

Axelrod, A. *The Evolution of Cooperation*. New York. Basic Books, 1984.

Hofstadter, D.R. *Metamagical Themas*. New York. Basic Books, 1985.

Luce, R.D. & Raiffa, H., *Games and Decisions*. New York, Dover, 1985.

Narendra, K. & Thathatchar, M. *Learning Automata: An Introduction*. New York. Addison-Wesley, 1989.

Rapoport, A. *Two Person Game Theory*. Ann Arbor. Un. of Michigan Press, 1966.