Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Games (CIG 2007)

Adaptation of Iterated Prisoner's Dilemma Strategies by Evolution and Learning

Han Yang Quek and Chi Keong Goh
Department of Electrical and Computer Engineering
National University of Singapore
{g0500073, ckgoh}@nus.edu.sg

Abstract—**This paper examines the performance and adaptability of evolutionary, learning and memetic strategies to different environment settings in the Iterated Prisoner's Dilemma (IPD). A memetic adaptation framework is devised for IPD strategies to exploit the complementary features of evolution and learning. In the paradigm, learning serves as a form of directed search to guide evolutionary strategies to attain good strategy traits while evolution helps to minimize disparity in performance between learning strategies. A cognitive double-loop incremental learning scheme (ILS) that encompasses a perception component, probabilistic revision of strategies and a feedback learning mechanism is also proposed and incorporated into evolution. Simulation results verify that the two techniques, when employed together, are able to complement each other's strengths and compensate each other's weaknesses, leading to the formation of good strategies that adapt and thrive well in complex, dynamic environments.**

Keywords: Evolution, Incremental Learning, Memetic Algorithm, Iterated Prisoner's Dilemma

I. Introduction

The Iterated Prisoner's Dilemma (IPD) is an abstract mathematical game where players cooperate or defect simultaneously without prior communication over repeated rounds. Though conceptually simple, IPD has been used to model behavioral developments and gain insights from social, political, economic and other interactions. In recent years, computational intelligence has made significant contributions to IPD. For instance, evolutionary algorithms (EAs) are well suited for searching robust strategies and analyzing IPD interactions. The core issue considered in this paper is the adaptation of strategies in various environments. Existing works show that EA are highly successful in discovering complex and effective methods of adaptation to very rich situations [1]. In particular, evolved strategies are very good at developing specialized adaptations to specific settings, capable of defending against defectors and cooperating with cooperators [2]. While existing works are mostly concerned with the generalization ability of evolved strategies [3], this paper focused on the adaptability of strategies to different environmental settings.

Learning is another adaptation framework that has been extensively studied in game theoretic problems [4], [5]. According to Hingston and Kendall [6], learning creates adaptive strategies that thrive in competitive settings by exploiting non-adaptive strategies. It is closely knitted to evolution and presents a similarly realistic scenario where knowledge accumulated through previous game play can be kept and used by players in the IPD tournament.

However, the pattern of decision making is rarely constant [7] but dependent on the environment and complex interaction between competing agents. It is known that evolution and learning have distinct advantages and disadvantages. Evolution facilitates information exchange among strategies but is limited by poor exploitation abilities. Generation of new individuals is not guided by lessons learned from past generations, but rather a form of trial and error process [8], often causing populations of IPD strategies to reach a naïve state when the algorithm terminates. On the contrary, learning allows the strategies to make complex decisions spontaneously but entails large score variance due to diverse learning experiences [9] across different players.

The paper considers development of a memetic adaptation framework [10] for strategies to exploit the complementary features of evolution and learning. In addition, a cognitive double-loop incremental learning scheme that encompasses a perception component, probabilistic revision of strategies and a double-loop learning mechanism is proposed. Integrating evolution with the cognitive learning scheme introduces a fair degree of realism to the behavioral modeling of players. A series of simulation is performed to uncover new insights into the intricacies between evolution and learning, justifying how adaptive strategies with good performance are created via memetic learning. Organization of the paper is as follows: Section II presents an overview of the IPD problem. Section III introduces different models of adapting IPD strategies. Section IV presents the proposed cognitive learning scheme and Section V highlights details of implementing the adaptation strategies. Section VI evaluates the performance of strategies via three distinct case studies. Section VII concludes the paper with a summary, with some discussions on the results and areas where future work can be embarked.

II. Iterated Prisoner's Dilemma

The IPD problem pertains to the study of short term rational decision of self-interest against the long term decision of overall interest. Each player has the option to COOPERATE (C) or DEFECT (D) and the payoff attained after each round is given by a Payoff Matrix as shown in Fig. 1. The game is played repeatedly among numerous strategies, each with its own set of characteristics and behaviors. According to Folk theorem, the set of Nash equilibriums of infinitely repeated rounds contains the cooperative solution [11]. Repeated defection or cooperation is not the best decision as strategies can perform much better by cooperating with reciprocal cooperators, exploiting unconditional cooperators, and resisting defectors [12].

		Player 2	
		COOPERATE	**DEFECT**
Player 1	**COOPERATE**	3,3	0,5
	DEFECT	5,0	1,1

Fig. 1: IPD Payoff Matrix

III. IPD ADAPTATION MODELS

A. Evolution

Evolution, as an optimization paradigm, is widely used to evolve strategies in the IPD game [1]. By retaining fit strategies and discarding weaker ones episodically, eventual convergence towards robust and effective strategies [3] is achieved. Many variants of evolutionary implementations have existed as of today. Evolving strategies are represented in binary forms [13], neural networks [14], real number [15] coded strings and even finite state machines (FSMs) [16]. Good strategies are selected in different ways as well, e.g. truncation selection [17], (μ, λ) and $(\mu + \lambda)$ selection [18] fitness-proportional selection [6]. Variation operators used also differs across implementations. While most EAs use a combination of crossover and mutation [13], pure mutation operators is used by Hingston and Kendall [6] and Chong and Yao [18] for their coevolutionary framework to analyze and study various aspects of the IPD problem. Other EAs also use speciation or niching to maintain genetic diversity [19] and elitism to avoid the lost of good strategies from the mating pool. Despite differences in the various implementations, the fundamental evolutionary framework is essentially the same and can be summarized as a sequential process of fitness assessment, genetic selection and genetic variation.

B. Learning

The learning methodology can be progressive [20] or reactionary [21]. Progressive learning schemes such as hill-climbers and gradient-based techniques are commonly applied to static environments where conditions are fixed and the notion of optimality is explicitly defined. Reactive learning is more applicable to a dynamic setting where the notion of optimality is changing or totally not in existence. Notable examples include the Pavlovian learning scheme and other stochastic searches. In its classical form, learning only affects the individual strategies with no facility for communication. Nonetheless, population-based learning in [13] was found to supersede its evolutionary counterpart. In general, learning functions as a local search operator that steers strategies in the direction which is deemed more "favorable" in the context of the game. It exploits domain information available at hand to improve performance of strategies based on some form of heuristics. Since the pattern of decision making is rarely constant [7] but highly dependent on the environment and complex interaction between competing agents, learning should be performed on an incremental basis, with partial memory [22] of past experiences. This scheme of learning better models the IPD players, who are capable of making complex decisions spontaneously from time to time, based on some memory of their past actions.

C. Memetic Learning

Memetic Learning [23] is a hybrid adaptation framework that unifies learning and evolution as one single entity. In IPD, the notion of evolving strategies memetically is less well studied compared to learning and evolution until recent years. Two distinct implementations that are widely used are the Baldwinian [24] and Lamarckian [25] adaptation models. In Baldwinian Evolution, offspring do not inherit any learned abilities directly from their parents but merely experienced an increased capacity to learn new skills [26]. Learning is performed after every evolutionary episode. In Lamarckian Evolution, however, learning precedes evolution and desired traits that are acquired by parents in the course of their lives are passed down directly to their offspring [27]. Despite differences, the two models are similar in spirit. As opposed to learning, memetic learning employs evolution as a tool to facilitate information exchange between learning strategies, allowing knowledge acquired through learning to be shared using evolutionary operators. Learning is used as a form of directed search to guide evolving strategies to attain eventual convergence towards good strategy traits. "Meta-Pavlov Learning" [28], is a good example of a Baldwinian based memetic learning strategy. In general, deterministic strategies are well-suited to fixed setups but not very robust to environmental modifications. Evolutionary and learning strategies are able overcome this limitation via adaptation. In particular, it is hypothesized that memetic strategies, which harness the synergy between learning and evolution, will be able to acquire significantly better performance.

IV. COGNITIVE LEARNING

A double-loop incremental learning scheme (ILS) that is adopted by all learning and memetic strategies is presented in this section. It is characterized by a cognitive framework which incorporates perception into the decision making process of strategies during the game play. ILS strategies can perceive the nature of opponents, conduct a probabilistic revision of strategies, and possess a double-loop learning mechanism that facilitates recovery from mistakes. Strategy revision is based on a notion of success and failure. In general, ILS breeds good strategies that can react and adapt well to different opponent strategies, and in the process maximize the overall payoff.

A. Identification of opponent strategies

A simple classification heuristics is formulated based on the correlation between the received payoffs and opponent's likelihood to defect and cooperate. Opponents are classified into three broad categories, strategies with a tendency to defect (Exploiters), reciprocate cooperation (Reciprocals) or cooperate unconditionally (Cooperators). The nature of each opponent is mapped out according to the sum of payoffs received in the first three rounds of game play. The range of scores (0-15) are divided into three equal intervals, each corresponding to a class of opponents as shown in Fig. 2. This classification process forms the basis to gain insight into the nature of unknown opponents so as to facilitate the

adoption of good strategies during subsequent game play.

SCORE RANGE	0 - 4	5 - 10	11 - 15
NATURE OF OPPONENT	EXPLOITERS	RECIPROCALS	COOPERATORS

Fig. 2: Classification of opponent strategies

B. Basis of strategy revision

The proposed learning scheme is built upon John Nash's [29] concept of Nash Equilibrium. It is conceptualized that each pair of strategies, despite their complexity and nature, will have a desired state at each round of game play, where both execute their best responses to each other. Each classification of opponent strategies will thus entail a desired response, defined by a Taxonomy Matrix in Fig. 3. The Matrix forms a heuristics that dictates the direction where local search is directed and is adhered to by all learning strategies throughout the course of game play.

		Player: COOPERATE	Player: DEFECT
Opponent	COOPERATE	Reciprocals	Cooperators
	DEFECT	-	Exploiters

Fig. 3: Taxonomy Matrix

To form the basis of learning, outcome of each round is classified as "success" (S) or "failure" (F). A "success" trial refers to an outcome where the player successfully attains the expected payoff against the classified opponent while a "failure" trial denotes otherwise. By keeping a record of the S and F counts accumulated by each strategy bit, strategy revision can take place. The S and F counts indicate how well a strategy fair against the opponent and are used to decide whether a strategy bit should be revised or left unchanged. The Taxonomy Matrix and underlying notion of success and failure are proposed to refine the Performance Matrix used by the Pavlovian learning scheme, which defines "success" as receiving Temptation (T) or Reward (R) payoffs and "failure" as awarded Punishment (P) or Sucker (S) payoffs. This is not a good way of defining the matrix due to the following:

1) Receiving P in the context of exploiters which defect perpetually should be considered good.
2) Receiving R in the context of unconditional cooperators should be considered bad as there are further opportunities for exploitation.
3) Receiving T in the context of reciprocals is not the best policy as it can well lead to endless cycles of retaliation.

In general, "success" and "failure" hold a fuzzy meaning when the payoff is P, R or T. With no knowledge about the opponent, uncertainty is involved during learning, as what is good for one opponent might be bad for another. Since "success" and "failure" decides the "goodness" of learning and exerts great impact on the performance of strategies, the proposed notion of "success" and "failure" is dynamically updated based on the perceived nature of the opponent.

C. Probabilistic revision of strategies

Probabilistic learning is used by the proposed adaptation scheme to revise strategy bits via the accumulated S and F counts over the entirety of game play. Fitter bits have larger S counts while weaker ones have larger F counts. Unused bits will contain a zero for both counts. An update (changing C to D or vice versa) is made only if the ratio of F counts to the total number of S and F counts exceeds a probability Ps. Mathematically,

$$\text{Swap(True) iff } F / (S + F) > Ps ,$$

Ps can be adjusted to suit the desired failure tolerance – amount of failure which a learning player is willing to undertake before deciding to revise his strategy. A higher Ps makes a player more tolerant to failures and less likely to revise its strategy. The number of games played using a bit, (S+F), before learning also impacts a player's sensitivity to environment changes. A large (S+F) delays learning but allows the "goodness" of a bit to be assessed via a wider observation window. An inherent tradeoff arises between the need to react promptly to changes in the pattern of opponent game play versus the need to maintain a sizable window of past experiences before performing strategy revision. Appropriate values of Ps and (S+F) are selected so that performance of the learning player is maximized. S and F are reset to zero whenever updating takes place and on encountering a new opponent. This prevents the past performance of a bit from affecting its current performance.

The above formulation ensures that desirable traits are likely to remain intact while undesirable ones are more susceptible to change. The inherent randomness regulates the learning pressure and avoids repeated updating of strategy bits when learning an incorrect move. It also takes into account the very fact that decision making is never absolute but subjected to a fair degree of uncertainty.

D. Double loop learning

The double-loop learning process draws parallel to human way of learning: perceiving, reasoning, self-evaluating and readjusting. A separate learning cycle which involves total reclassification of the opponent and re-mapping of the perceived best response is performed in situations where it is mapped incorrectly. Reclassification is triggered when the accumulated F count of any strategy bit used in the game play exceed a count of f. Notion of "success" and "failure" is changed and a new perceived best response adopted. If the new perceived best response corresponds to the desired one, the number of F counts will be greatly reduced, indicating that the player is using the right tactics against the opponent. Otherwise, learning continues until the perceived and desired best response coincide.

While inner-loop learning gives players the ability to form models of mental perception about opponents and learn incrementally, outer-loop learning allows players to perform evaluation of each model and make appropriate readjustments from time to time. By learning and relearning [30], strategies can adapt and realign their tactics dynamically to changes in the nature of unknown opponents, via formation, evaluation and revision of perceptions. A flow chart of the double-loop learning scheme is shown in Fig. 4. Unlike absolute reactionary learning, the integrated double-loop learning prevents chances of entering a loop of

endless updating and lowers the possibility of being trapped in a local minimum.

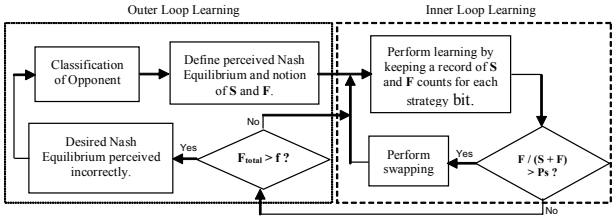


Fig. 4: Overview of the Double-Loop Learning Process

V. IMPLEMENTATION

Evolution of strategies is achieved by means of a genetic algorithm (GA) model as described in section 3. Learning is performed by the ILS in section 4 while the memetic algorithm (MA) combines evolution and ILS based on the framework in section 3. The flowchart of MA is presented as shown in Fig. 5.

Each GA player is represented by a 65-bit chromosome. The first bit encodes the initial condition for triggering the first move at the start of each game set while the remaining bits address the 64 (2^6) possible histories of 6-bit memory configurations corresponding to the previous three moves of both players. Each ILS and MA player is represented by a 65 by 3 2D array where each slot comprises of three genes. The first gene denotes the action corresponding to a 6-bit memory configuration while the second and third store the S and F counts. All strategies also encode an independent 6-bit memory that provides information about round histories and is used to decide the next move. The initial population of all adaptation strategies is randomly generated. In GA and MA, fitness assignment is performed after each tournament and is given by the payoffs accumulated throughout the game play. Niching with a sharing distance of **r** is applied to encourage the growth of a diverse repertoire of strategies. Elitism is incorporated by means of binary tournament selection. Selected individuals will then undergo uniform crossover and binary uniform mutation.

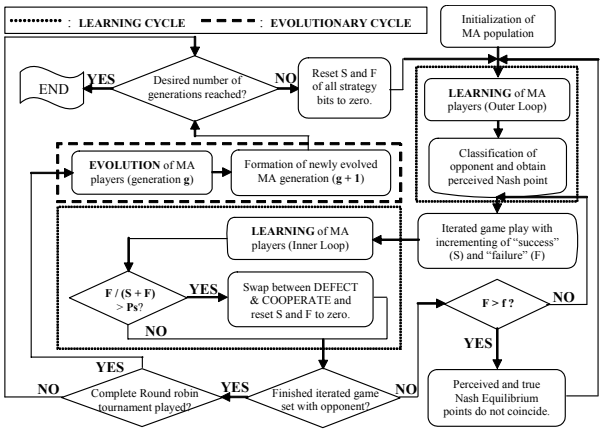


Fig. 5: Simple flowchart depicting the operations of MA

VI. SIMULATION RESULTS

Simulations are carried out using the Visual C++ development kit. A summary of the parameter values used in the simulations are depicted in Table I.

TABLE I
LIST OF PARAMETER VALUES USED IN THE SIMULATION RUNS

Parameters	Values
No. of rounds in an iterated game set, **α**	200
Generation, **g**	600
Population size, **n**	Variable
Number of strategies for each population type, **p**	30
Tournament size, **s**	2
Crossover rate, **c**	0.8
Mutation rate, **m**	0.05
Niche radius, **r**	50
Failure count, **f**	10

A. Case Study 1: Performance Assessment

The first case study compares the relative performance of GA, ILS and MA when each competes with AllC, AllD and TFT. The base strategies contain a good mix of cooperators, defectors and reciprocals, each with its own unique best response. Three tests, A, B and C are used to evaluate the performance of GA, ILS and MA against single, multiple opponents as well as against one another. Group performance will be analyzed using generation payoffs, box plots, ideal player scores (IPS) and statistical tests.

Test A – Single opponent type

Each of the three adaptation strategies is set to compete against AllC, AllD, TFT and their own players separately. The IPS - average payoff attained against the maximum attainable payoff per game, of GA, ILS and MA is plotted against the ideal IPS of 5, 3, 3, 1 in Fig. 6 when playing against AllC, TFT, itself and AllD respectively.

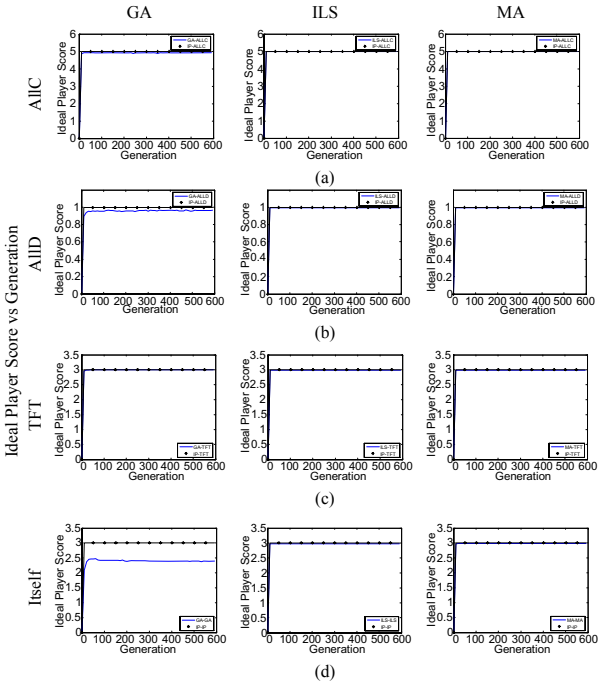


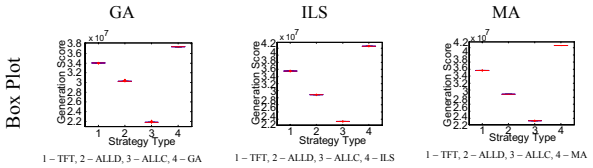
Fig. 6: IPS against (a) AllC, (b) AllD, (c) TFT and (d) itself

The plots show that all three strategies can achieve IPS

close to the ideal IPS for single opponent type. In particular, IPS of ILS and MA is almost indistinguishable from the ideal IPS. For GA, greater deviation from the ideal IPS reinforces the fact that unguided evolution is limited in its ability to track the desired best response. Onset of premature convergence places a limit on GA's ability to evolve players that can cooperate well with one another in Fig. 6(d). ILS and MA players can perform significantly better by virtue of their ability to adjust strategies constantly in the search for the desired best response.

Test B – Multiple opponent types

Extending from test A, each adaptation strategy is set to play against AllC, AllD and TFT simultaneously. Box plots in Fig. 7 show that GA, ILS and MA are still the best among each set. In particular, ILS and MA outperform GA significantly by virtue of the ability to cooperate with TFT and defect against AllC and AllD to a larger extent.

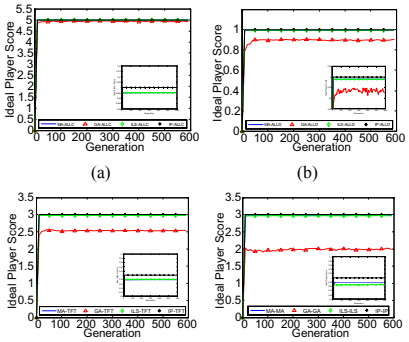


(a) (b) (c)

Fig. 7: Box plots of (a) GA, (b) ILS and (c) MA

Compared to the single opponent case, plots in Fig. 8 also show that the IPS of GA has degraded a lot in the face of multiple opponent types. Since evolution is performed only after each tournament, GA players do not have the luxury to alter their strategies within the course of game play. The evolved strategies are essentially fixed and tradeoffs are involved when attempting to balance between different best responses, as what is good against an opponent is not necessarily good for another. Instead of achieving the best performance, GA, at best, can only score reasonably well against each opponent type. ILS and MA can perform better by adjusting to changes in the nature of opponent. Subtle difference in Fig 8(d) also suggests that MA players are more cooperative among themselves. Further statistics in Table II also show that MA and ILS perform much closer to the ideal player than GA. MA is also superior to ILS in terms of the score uniformity across players and inference can be made to point out the understated fact that evolution with learning can stabilize performance across a population, especially when players have differing learning experiences.

Analysis of the learning traces of ILS and MA in Fig. 9 also reveal that learning is the dominant force in the short run and is responsible for the initial improvement of players. Evolution exerts pressure for improvement in the long term basis when the learning ratio is relatively low. The lower and less fluctuating MA trace indicates that MA players need to learn less on an individual basis due to the possibility of information exchange via evolution. Competent strategies will eventually be adopted by weaker players, affirming that evolution is capable of leveling out differences in learning experience between players and stabilize scores across an entire population of strategies.



(a) (b) (c) (d)

Fig. 8: IPS against (a) AllC, (b) AllD, (c) TFT and (d) itself

TABLE II
IPS ERROR AGAINST (A) ALLC (B) ALLD, (C) TFT & (D) ITSELF

	GA	ILS	MA
Mean	60.5405	6.5370	6.0160
Std	50.5148	0.2587	0.0081
Median	52.7405	6.5458	6.0137
25th Percentile	9.8904	6.4967	6.0112
75th Percentile	95.5872	6.7105	6.0194

(a)

	GA	ILS	MA
Mean	63.7098	3.0596	3.0115
Std	93.0117	0.0273	0.0022
Median	10.6724	3.0631	3.0110
25th Percentile	7.7254	3.0572	3.0098
75th Percentile	115.2993	3.0774	3.0129

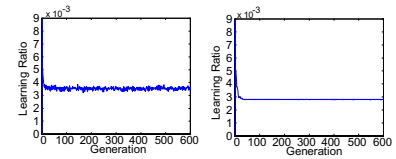
(b)

	GA	ILS	MA
Mean	261.0971	12.2445	12.7109
Std	104.9512	0.1003	0.0293
Median	253.2133	12.2497	12.7079
25th Percentile	187.0430	12.2272	12.6865
75th Percentile	304.9748	12.3266	12.7265

(c)

	GA	ILS	MA
Mean	621.7026	20.0730	12.4371
Std	146.6414	3.7954	0.1981
Median	590.8009	20.1469	12.3905
25th Percentile	506.9482	19.7961	12.2863
75th Percentile	731.0408	22.4862	12.5023

(d)



(a) (b)

Fig. 9: Learning traces for (a) ILS and (b) MA against multiple opponents

Test C – Relative performance of GA, ILS and MA

To assess the overall strategic dominance, GA, ILS and MA are set to compete against one another in the presence of AllC, AllD and TFT. Despite small IPS deviations in Fig. 10, ILS and MA are still able to track the ideal IPS fairly well. Nevertheless, the IPS of GA is worst off because the addition of adaptive opponents has upset the equilibrium state of the evolved GA strategies and their ability to balance between the distinct best responses. It can also be deduced from the persistently fluctuating IPS profile that there is neither one overall optimal strategy nor equilibrium state for GA to converge to.

The normalized generation scores in Fig. 11(a) show that ILS and MA continue to lead other strategies by a large score margin. Drop in performance of GA below TFT is possibly due to the exploitation by ILS and MA, but in greater likelihood, the inability of each fixed, evolved GA strategy to cope with the added complexity in the setup. As much ease as it seems with only six strategies, addition of adaptive opponents entails multiple best responses that are constantly shifting. The dynamics involved has become insurmountable for any GA strategy to handle. This asserts that the inherent ability to learn dynamically is an important adaptation trait for preserving the performance of strategies amidst changes in complexity over time and across varying environments. Box plot in Fig. 11(b) also shows that MA outperforms ILS by a diminutive amount.

To justify that this lead is significant, the score difference across all strategies is verified via the Kolmogorov-Smirnov-Test (KS-Test) where the maximum difference between the cumulative score distribution functions of two strategy types, over all possible player scores, is used as its test statistic. The p-value of the test will decide how different two populations of strategies are. A p-value close to 0 denotes distinct difference while a p-value close to 1 denotes close correlation. Statistical tabulation of p-values between GA, ILS, MA and all strategies in Table III show explicitly that GA, ILS and MA are significantly different from strategies other than itself. Coupled with observations attained previously, it can be concluded that the superiority of MA over ILS is indeed noteworthy and non-trivial.

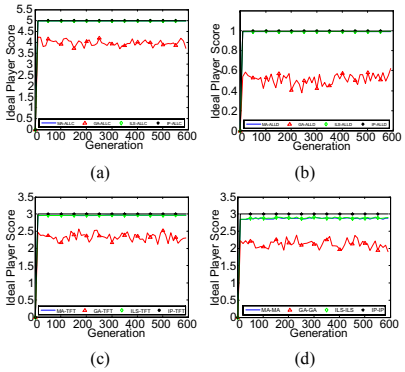


Fig. 10: IPS against (a) AllC, (b) AllD, (c) TFT and (d) itself

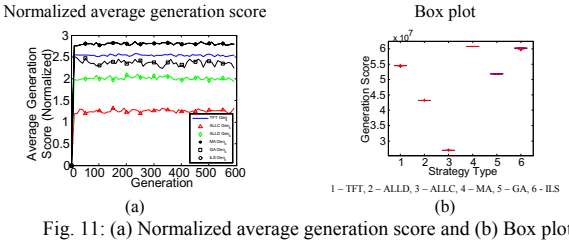


Fig. 11: (a) Normalized average generation score and (b) Box plot

TABLE III
P-VALUES OF KS-TEST BETWEEN STRATEGIES

	MA	GA	ILS	ALLC	ALLD	TFT
MA	1	1.7973e-014	1.7973e-014	1.7973e-014	1.7973e-014	1.7973e-014
GA	1.7973e-014	1	1.7973e-014	1.7973e-014	1.7973e-014	1.7973e-014
ILS	1.7973e-014	1.7973e-014	1	1.7973e-014	1.7973e-014	1.7973e-014

B. Performance in different environments

Case study 2 is devised to show that MA leads to better performance and adaptation on a broader perspective. The adaptability of GA, ILS and MA is assessed with different mixture of opponent strategies. Pavlov, RAND, TFTT and STFT are added to the existing base strategies in various combinations via four different test sets. All plots in various test sets show that the dynamism in learning still allows MA and ILS to maintain substantial lead over all strategies. Fig. 12(a) also justifies that the proposed cognitive learning scheme indeed offers better performance than Pavlov. Onset of large-scale fluctuation suggests that GA's ability to score well against all opponents is compromised in the midst of trying to cope in the complex environment. Nonetheless, the score uniformity in GA suggests that evolution is important to smooth out disparity across players' scores in different environments. Fig. 12(b)-(d) show that the dynamic double loop learning allows ILS and MA to eventually work towards mutual cooperation with reciprocals and defection against exploiters. The inability to cooperate or defect fully in the midst of balancing multiple tradeoffs and goals only allows GA to perform reasonably well but not surpass other strategies significantly. In the extreme case, GA actually evolves into RAND by virtue of the complexity in Fig. 12(d). This indicates that it is practically impossible to define any fixed pattern of game play since each opponent differs in nature according to its own right. The KS-test results of all test sets also demonstrate that MA's payoff is distinctly the best among all strategies, including its close competitor ILS. Synergy between evolution and learning has enabled MA to sustain its lead consistently throughout all test sets. On the other hand, ILS, in the absence of evolutionary pressure, suffers large differential in payoffs, which inevitably brings down the population performance.

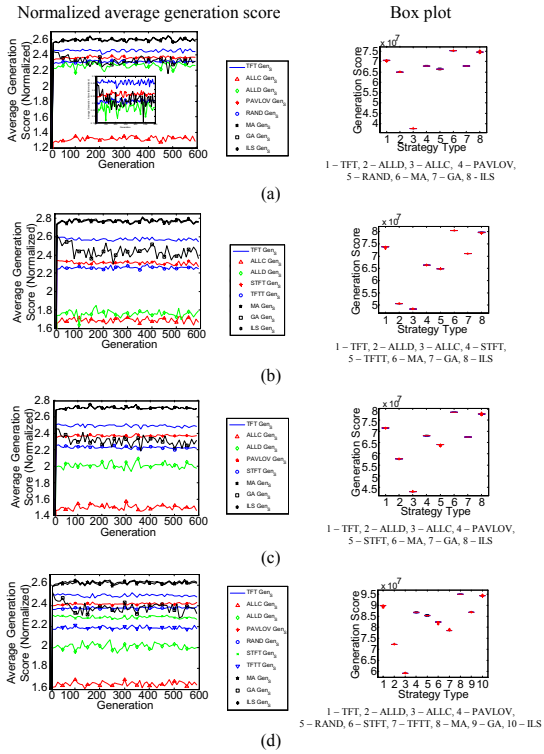


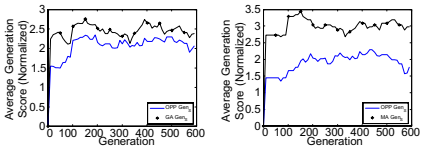
Fig. 12: (a) Normalized average generation score and (b) Box plot

C. Performance in dynamic environments

The final case study evaluates the performance of MA and GA in a dynamically changing environment. With the opponent changing probabilistically to AllC, AllD or TFT after every 50-100 generations, it is justified from the large score margin in Fig. 13 that MA performs better than GA. The increase in MA's score comes at the expense of the opponent, indicating MA's ability to exploit the opponent when GA failed to do so. While GA converges to a strategy that is generally cooperative, MA is able to go one step further to reap the temptation payoffs in the case of AllC and to fend against AllD as well. Collectively, this leads to

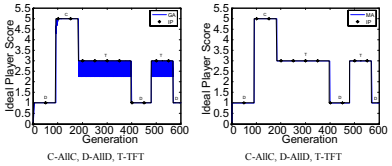
an improvement in the generation score and affirms the good adaptive ability of MA over GA.

To bring the discussion further, results from 2 distinct runs are used to assess the IPS over 600 generations. Fig. 14(a) shows discernible details that GA is plagued by large scale fluctuation when the opponent transits from AllC or AllD to TFT. Initially, GA has been so accustomed to defect against AllC and AllD that it is able to track the ideal IPS with negligible error. Nonetheless, sudden transition to TFT radically changes defection from good to totally bad. Convergence of GA towards defection and the limited variation ability to adopt cooperative traits prevents GA from breaking out of the instability zone. This instability is however not experienced by MA in Fig. 14(b) because during instances when the opponent transits between two radically different types, players will experience a sudden increase in F counts and this triggers reclassification when the threshold of change is reached. MA players are thus able to break out of their old mental model and reshape their perception of the opponent. Compared to random mutation, learning makes strategy revision more explicit. As long as a portion of the population has perceived the right best response, information is propagated to other members via evolution and traits are adjusted almost immediately. Fluctuation is minimized and perfect tracking of the ideal IPS is achieved. With more transitional phases in Fig. 15, results also show that MA adapts well to the changing nature of opponents on a consistent basis.



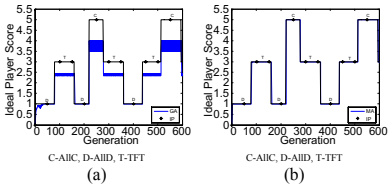
(a) (b)

Fig. 13: Normalized average generation score for (a) GA and (b) MA



(a) (b)

Fig. 14: IPS of (a) GA and (b) MA against ideal player in run 2

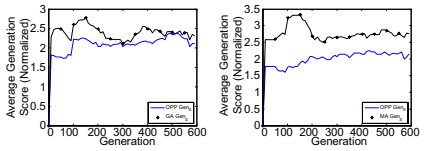


(a) (b)

Fig. 15: IPS of (a) GA and (b) MA against ideal player in run 8

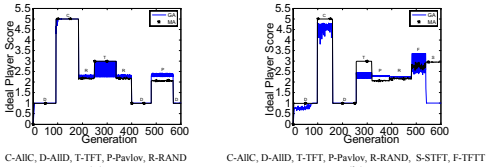
Finally, randomness of the environment is enhanced by allowing opponents to morph into AllC, AllD, TFT, Pavlov, STFT, TFTT or RAND over the same frequency of change as introduced previously. It is apparent from Fig. 16 that GA's score has dropped marginally while the opponent score has improved significantly. Difficulty of the opponent has risen considerably, given the fact that it can now resume a more probabilistic nature apart from deterministic ones. Decline in score is due to the inability to track the desired best responses with certainty. Nonetheless, MA players are still able to secure a large score advantage by virtue of the already huge score margin as attained previously.

Results from 2 separate runs are again used to compare the IPS over 600 generations. Since it is difficult to set an ideal IPS for probabilistic players, IPS of GA and MA are instead superimposed on the same plot for easy comparison. Fig. 17 shows that GA is still plagued by large and small scale fluctuation. Large scale ones occur when the opponent behaves as TFT while small ones subsist when the opponent assumes a probabilistic outlook. A probable explanation is because the evolved GA strategy has evolved to play well against opponents in the initial phases of evolution. Subsequently, the low mutation rate limits GA's ability to create perfectly cooperative individuals from defect-oriented genotypes e.g. large number of bit flips from 0s to 1s is needed. Mixed strategies with mild chances of cooperation are formed instead. Alternate cooperation and defection against TFT predominantly sets the fluctuation profile into play. Fluctuation occurs with a lesser extent for probabilistic opponents since it is harder to coin whether GA players will perform better if they were cooperative or defect-oriented, unlike the case of TFT where it is clear that cooperation will yield the best payoffs. Comparatively, with a common framework of double loop learning, fluctuation is milder for MA players. Radical change in learning direction during outer loop learning, coupled with the high frequency of bit revision during inner loop learning, is able to introduce substantial change of strategy traits even if MA players have converged to be significantly similar under evolution. Learning allows players to adopt strategies that are vastly different from those used against previous opponents and thus reduces the performance dependency on the nature of opponents. This allows the entire population to adapt and adjust smoothly when the opponent transits between two different strategy types. Though MA population suffers a small degree of fluctuation against probabilistic opponents, this is notably less compared to the GA population.



(a) (b)

Fig. 16: Normalized average generation score for (a) GA and (b) MA



(a) (b)

Fig. 17: Average score of GA plotted against MA in (a) run 2 and (b) 13

VII. CONCLUSION

In this paper, the performance and adaptability of evolutionary, learning and memetic strategies are assessed in various IPD settings. A memetic adaptation framework, MA, is developed to harness the synergy between evolution and learning. In this framework, learning assists evolutionary strategies to acquire good strategy traits and to react spontaneously to changes in the environment while evolution provides an avenue to minimize performance disparity between learning players via knowledge exchange. A cognitive double-loop incremental learning scheme, ILS, which encompasses a perception component, probabilistic revision of strategies and a double-loop learning mechanism is also proposed and incorporated into the evolutionary process to correct some flaws in Pavlovian Learning and to model IPD players more realistically.

Comparative study conducted for different environment setups showed that players adapted by MA are superior in performance to GA and ILS. GA suffers from instability and deteriorating performance when multiple opponent strategy types are introduced while ILS suffers from diverse learning experiences among individuals, leading to large score variance which undermines the performance of the entire population. On the contrary, the combination of incremental learning and evolution in MA allows players to balance the task of exploration and exploitation of different strategies while preserving the trend of dominance consistently throughout different settings. It is gathered that both learning and evolution are essential elements in the IPD game. Their concurrent interaction is crucial for the formation of good strategies that adapt and thrive well in complex, dynamic environments. Future work can include simulation of other strategies, formulating better learning schemes, applying memetic learning to noisy settings, devising complex payoff matrices and conducting evolutionary tournaments. A thorough study of more complicated situations like the above-mentioned would be useful in giving greater insight to the intricacies and complexity involved in the IPD game.

REFERENCES

[1] Axelrod, R., "The evolution of strategies in the iterated prisoner's dilemma," in Lawrence Davis (ed.) *Genetic Algorithm and Simulated Annealing*, London: Pitman, 1987, pp. 32-41.

[2] Axelrod, R., *The Complexity of Cooperation*. NJ: Princeton University Press, 1997.

[3] Darwen, P. and Yao, X., "On evolving robust strategies for the iterated prisoner's dilemma," in *Progress in Evolutionary Computation*, ser. Lecture Notes in Artificial Intelligence, 1995, vol. 956, pp. 276-292.

[4] Fudenberg, D. and Levine, D. K., *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[5] Thrun, S., "Learning to play the round of chess," *Machine Learning*, vol. 40, issue 3, pp. 243-263, September 2000.

[6] Hingston, P. and Kendall, G., "Learning versus evolution in iterated prisoner's dilemma," in *Proceedings of the Congress on Evolutionary Computation 2004 (CEC'04)*, Portland, Oregon, vol. 1, June 2004, pp. 364-372.

[7] Eoyang, G., *Genetic Algorithm as Decision Support Tool*. Chaos Limited, 1996.

[8] Michalski, R. S., "Learnable evolution model: evolutionary processes guided by machine learning," *Machine Learning*, vol. 38, issue 1-2, pp. 9-40, February 2000.

[9] Lamma, E., Riguzzi, F. and Pereira, L. M., "Belief revision via Lamarckian evolution," *New Generation Computing*, vol. 21, no.3, pp. 247-275, May 2003.

[10] Carrier, C. G., "Unifying learning with evolution through Baldwinian evolution and Lamarckism: a case study," in *Proceedings of the Symposium on Computational Intelligence and Learning (CoIL-2000)*, June 2000. pp. 36-41.

[11] Bodo, P., "In-class simulations of the iterated prisoner's dilemma round," *Journal of Economic Education,* vol. 33, no. 3, pp. 207-216, 2002.

[12] Neill, D. B., "Optimality under noise: higher memory strategies for the alternating prisoner's dilemma," *Journal of Theoretical Biology,* vol. 211, pp. 159–180, 2001.

[13] Gosling, T., Jin, N. and Tsang, E., "Population based incremental learning versus genetic algorithms: iterated prisoners dilemma," Department of Computer Science, University of Essex, England, Tech. Rep. CSM-401, March 2004.

[14] Chong, S. Y. and Yao, X., "The Impact of Noise on Iterated Prisoner's Dilemma with Multiple Levels of Cooperation," in *Proceedings of the Congress on Evolutionary Computation 2004 (CEC'04)*, Portland, Oregon, vol. 1, June 2004, pp. 348-355.

[15] Hales, D., "Change Your Tags Fast! - A Necessary Condition for Cooperation?," in *Proceedings of the Joint Workshop on Multi-Agent and Multi-Agent-Based Simulation*, July 2004, pp. 89-98.

[16] Fogel, D. B., "On the relationship between the duration of an encounter and the evolution of cooperation in the iterated prisoner's dilemma," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 3, pp. 349-363, 1996.

[17] Fogel, D.B., Fogel, G. B. and Andrews P. C., "On the instability of evolutionary stable states," *Biosystems*, vol. 44, pp. 135-152, 1997.

[18] Chong, S. Y. and Yao, X., "Behavioral Diversity, Choices and Noise in the Iterated Prisoner's Dilemma," *IEEE Transactions on Evolutionary Computation,* vol. 9, no. 6, pp. 540-551, 2005.

[19] Darwen, P. and Yao, X., "Speciation as automatic categorical modularization," *IEEE Transactions on Evolutionary Computation,* vol. 1, no. 2, pp. 101-108, 1997.

[20] Salles, B., "Constructing progressive learning routes through qualitative simulation models in ecology," in *Proceedings of the International Workshop on Qualitative Reasoning, QR'01*, May 2001, pp. 82-89.

[21] Lin, L. J., "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine Learning*, vol. 8, issue 3-4, pp. 293-321, May 1992.

[22] Maloof, M. A. and Michalski, R. S., "Incremental learning with partial instance memory," in *Proceedings of the 13th International Symposium on Foundations of Intelligent Systems,* 2002, pp. 16-27.

[23] Moscato, P., "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms," California Institute of Technology, Pasadena, California, USA, Tech. Rep. Caltech Concurrent Computation Program, Report. 826, 1989.

[24] Baldwin, J. M., "A new factor in evolution," American Naturalist vol. 30, pp. 441-451,536-553, 1896.

[25] Ong, Y. S. and Keane, A. J., "Meta-Lamarckian learning in memetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 8, no.2, pp. 99-110, 2004.

[26] Deacon T., *The Symbolic Species: the Coevolution of language and human brain*. London: Penguin, 1997.

[27] Sasaki, T., and Tokoro, M., "Adaptation towards changing environments: Why Darwinian in nature?" in Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97), Brighton, UK, 28–31 July*, 1997, pp. 145–153 Cambridge, MA. MIT Press / Bradford Books, Cambridge, MA.

[28] Suzuki, R. and Arita, T., "Interactions between Learning and Evolution: Outstanding Strategy generated by the Baldwin Effect," *Biosystems*, vol. 77, issue 1-3, pp. 57-71, 2004.

[29] Nash, J. F. Jr., "Equilibrium Points in n-Person Games," *in Proceedings of the National Academy of Sciences 36,* U.S.A., 1950, pp. 48-49.

[30] Liang, K. H., Yao, X. and Newton, C., "Lamarckian evolution in global optimization," in *Proceedings of the IEEE 26th Annual Conference of Industrial Electronics 2000, (IECON 2000)*, vol. 4, October 2000, pp. 2975-2980.