How Learning Can Affect the Course of Evolution in Dynamic Environments

Reiji SUZUKI Takaya ARITA
Graduate School of Human Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
E-mail: {reiji, ari}@info.human.nagoya-u.ac.jp

Abstract

The Baldwin effect is known as interactions between learning and evolution, which suggests that individual lifetime learning can influence the course of evolution without the Lamarckian mechanism. Our concern is to consider the Baldwin effect in dynamic environments, especially when there is no explicit optimal solution through generations and it depends only on interactions among agents. We adopt the iterated Prisoner's Dilemma as a dynamic environment and introduce phenotypic plasticity to strategies by using a meta-learning rule termed "Meta-Pavlov". In this simulation, the Baldwin effect was observed as follows: First, strategies with enough plasticity spread, which caused a shift from defective population to cooperative. Second, these strategies were replaced by the strategy [x00x], which has a modest amount of plasticity.

Keywords: baldwin effect, learning and evolution, the iterated prisoner's dilemma, artificial life.

1 Introduction

There have been a lot of discussions about interactions between learning and evolution. The Baldwin effect [1] is one of them, which suggests that individual lifetime learning can influence the course of evolution without the Lamarckian mechanism. This effect has come to the attention recently not only of biologists, but also of the computer scientists with the evolutionary simulation of Hinton and Nowlan [2]. Since Hinton and Nowlan, many studies have been conducted, most of which have discussed the effect on the assumption that environments are static and the optimal solution is fixed.

However, as we see in the real world, learning could be more effective and utilized in dynamic environments, because the flexibility of plasticity itself is advantageous to adapt ourselves to the changing world. Therefore, it is very important to examine how learning can affect the course of evolution in dynamic environments.

Our objective is to clarify the function and the mechanism of the Baldwin effect in dynamic environments focusing on balances between benefit and cost of learning, while most of the studies concerning the Baldwin effect have aimed at the static environments. In general, dynamic environments can be divided typically into the following two types: the environments in which the optimal solution is changed as the environment changes, and the ones in which each agent's fitness is decided by interactions with other agents.

As the former type of environments, Anderson [3] quantitatively analyzed how learning affects evolutionary process in the dynamic environment whose optimal solution changes through generations. Sasaki and Tokoro [4] studied the relationship between learning and evolution using a simple model, where individuals learn to distinguish poison and food by modifying the connective weights of neural network. These studies emphasized the importance of learning in dynamic environments.

We adopted the iterated Prisoner's Dilemma (IPD) as the latter type of environments, where there is no explicit optimal solution through generations and fitness of agents depends only on interactions among them. This paper describes the Baldwin effect briefly, explains our evolutionary model and discusses how this effect was observed in the evolutionary experiments.

2 Background

The Baldwin effect explains interactions between learning and evolution by paying attention to balances between benefit and cost of learning. The Baldwin effect consists of the following two steps (Turney, Whitley and Anderson [5]): In the first step, lifetime learning (phenotypic plasticity) gives individual agents chances to change their phenotypes. If the learned traits are useful for agents and make their fitness increase, they will spread in the next population. In the second step, if the environment is sufficiently stable, the evolutionary path finds innate traits that can replace learned traits, because of the cost of learning.

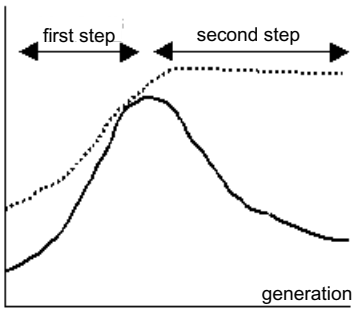

Figure 1: Two steps of the Baldwin Effect.

Through these steps, learning can accelerate the genetic acquisition of learned traits without the Lamarckian mechanism in general. Figure 1 shows the concept of the Baldwin effect roughly which consists of two steps described above.

We have adopted the iterated Prisoner's Dilemma (IPD) as a dynamic environment, which represents an abstraction of the situations causing social dilemma. Each "game" is carried out as follows:

- Two players independently choose actions from Cooperate (C) or Defect (D) without knowing the other's choice. Each player gets the score according to the payoff matrix (Table 1). We term this procedure "round".
- Above round is executed repeatedly and players compete for higher average scores.

In the only one round game, the payoff matrix makes defecting be the only dominant strategy regardless of opponent's action, and defect-defect action pair is the only Nash equilibrium. But this equilibrium is not Pareto optimal because the score of each player of cooperate-cooperate action pair is higher, which causes a dilemma. Furthermore, when the round is repeated, cooperating each other will turn out advantageous to both of them in the long run.

Table 1: Payoff matrix of Prisoner's Dilemma.

player \ opponent	cooperate	defect
cooperate	(R:3, R:3)	(S:0, T:5)
defect	(T:5, S:0)	(P:1, P:1)

(player's score, opponent's score)

$T > R > P > S, 2R > T + S$

3 The Model

3.1 Expression of strategies

The strategies of agents are expressed by two types of genes: genes for strategy (GS) and genes for learning (GL). GS describes deterministic strategies of IPD like Lindgren's model [6], which defines next action according to the history of actions. GL expresses whether each corresponding bit of GS is plastic or not.

A strategy of memory m has an action history h_m which is a m-length binary string as follows:

$$h_m = (a_{m-1}, \ldots, a_1, a_0)_2, \tag{1}$$

where a_0 is the opponent's previous action ("0" represents defection and "1" represents cooperation), a_1 is the previous player's action, a_2 is the opponent's next to previous action, and so on.

GS for a strategy of memory m can be expressed by associating an action A_k (0 or 1) with each history k as follows:

$$GS = [A_0 A_1 \cdots A_{n-1}] \quad (n = 2^m). \tag{2}$$

In GL, L_x specifies whether each phenotype of A_x is plastic (1) or not (0). Thus, GL can be expressed as follows:

$$GL = [L_0 L_1 \cdots L_{n-1}]. \tag{3}$$

For example, the famous Tit for Tat strategy (cooperates on the first round, whatever its opponent did on the previous round) can be described by memory 2 as GS=[0101], GL=[0000].

3.2 Meta-Pavlov learning

A plastic phenotype can be changed by learning during game. We have adopted a simple learning rule termed "Meta-Pavlov" and each agent changes phenotypes according to the result of each round by referring to the Meta-Pavlov learning matrix (Table 2). It doesn't express the strategy itself but the way to change own strategy (phenotype) according to the result of the current round, though this matrix is the same as that of Pavlov strategy (Nowak and Sigmund [7]). The learning process is described as follows:

- At the beginning of the game, each agent has the same phenotype as GS itself.
- If the phenotype used in the round is plastic (the bit of GL corresponding to the phenotype is 1), the phenotype is changed to the corresponding value in the Meta-Pavlov learning matrix based on the result of the round.

Table 2: Meta-Pavlov learning matrix.

player \ opponent	cooperate	defect
cooperate	C	D
defect	D	C

- The agent uses the new strategy specified by the changed phenotype from next round on.

Take a strategy of memory 2 expressed by GS=[0001] and GL=[0011], for example of learning (Figure 2). Each phenotype represents the next action corresponding to the history of the previous round, and the underlined phenotypes are plastic.

Let us suppose that the history of previous round was "CC (player's action: cooperation, opponent's action: cooperation)" and the opponent defects at the present round. This strategy cooperates according to the phenotype and the result of the present round is "CD". The strategy changes own phenotype based on the Meta-Pavlov learning matrix according to the result of the present round, because the phenotype applied at this round is plastic. The phenotype "C" corresponding to the history "CC" is changed to "D" in this example. Therefore, this strategy chooses defection when it has the history "CC" at the next time.

The values of GS that are plastic act merely as the initial values of phenotype. Thus we represent strategies by GS with plastic genes replaced by "x" (e.g. GS=[1000], GL=[1001] → [x00x]).

3.3 Evolution

Each bit of gene is set randomly in the initial population. The round robin tournament is conducted between individuals with the strategies as are described in the previous section, under the condition in which the performed action can be changed by the noise with probability p_n. Each plastic phenotype is reset to the corresponding value of GS at the beginning of games. Rounds are repeated with the probability p_f, which is decided at the end of each round. The tournament is "ecological": The total score of each agent is regarded as a fitness value, new population is generated by the "roulette wheel selection" according to the scores, and mutation is performed on a bit-by-bit basis with probability p_m.

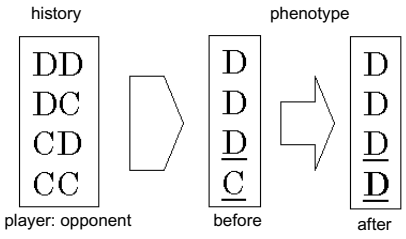

Figure 2: Example of Meta-Pavlov learning.

Average scores during the first 20 IPD games between new pair are stored, and will be used as the results of the games instead of doing games actually, so as to reduce the amount of computation. Stored scores are cleared and computed again by doing games every 500 generation.

4 Preliminary experiments

Strategies of memory 2 were investigated in the preliminary experiments. We conducted an evolutionary experiment for 2000 generations using the following parameters: $population = 1000$, $p_m = 1/1500$, $p_n = 1/25$ and $p_f = 99/100$.

The results are shown in Figure 3 and 4. The horizontal axis represents the generations. The vertical axis represents the distribution of strategies. At the same time, it also represents both "plasticity of population" (in black line) which is the ratio of "1" in all GLs and the average score (in white line). Plasticity of population is supposed to correspond to the "degree of dependence on learning" in Figure 1. The average score represents the degree of cooperation in the population, and it takes 3.0 as the maximum value when all rounds are "CC".

The evolutionary phenomena observed in a typical experiment are summarized as follows. Defective strategies ([0000], [000x] and so on) spread and made the average score decrease until about 60th generation, because these strategies can't cooperate each other. Simultaneously, partially plastic strategies ([0x0x], [00xx], [0xxx]) occupied the population. Next, around the 250th generation, more plastic strategies ([xxxx], [x0xx] and so on) established cooperation quickly, which made the plasticity and average fitness increase sharply. This transition is regarded as the first step of the Baldwin effect.

Subsequently, the plasticity of population decreased and then converged to 0.5 while keeping the average score high. Finally, the strategy [x00x] occupied the most of the population. The reason seems to be that the strategy has the necessary and adequate amount of plasticity to maintain cooperative relationships and prevent other strategies from invading in the population. This transition is regarded as the second step of

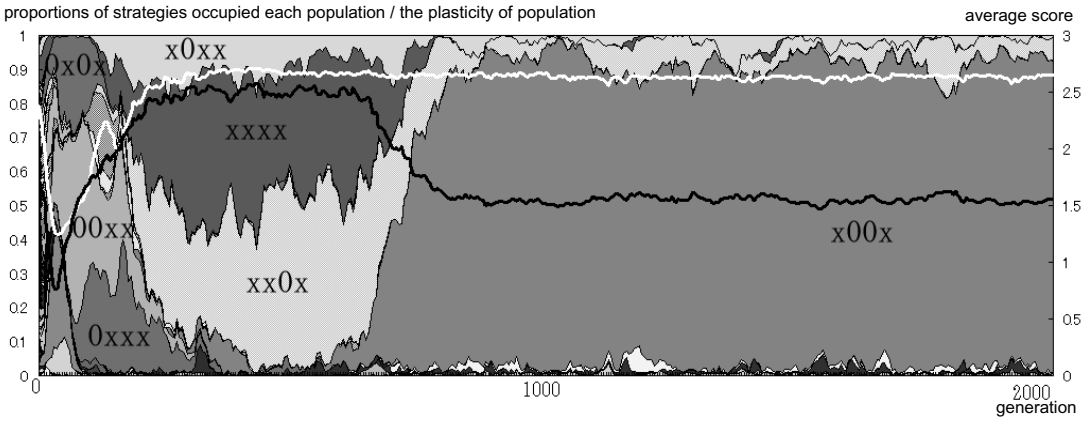

Figure 3: Experimental result (2000 generations).

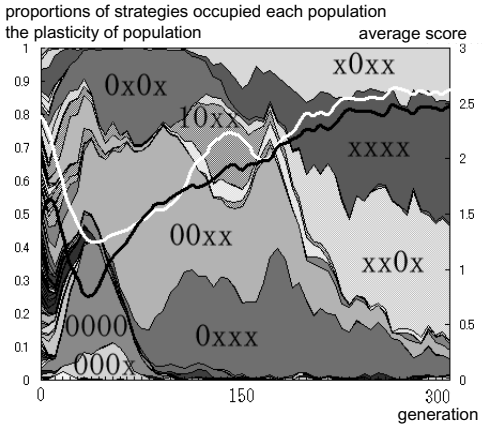

Figure 4: Experimental result (300 generations).

the Baldwin effect.

The population converged to the strategy [x00x] in all experiments, and the evolutionary phenomena described above was observed in 70% of experiments. Further analysis has shown that this strategy satisfies ESS condition in memory 2 population.

5 Conclusion

We have discussed how learning can affect the course of evolution in dynamic environments based on the results of the computational experiments on the evolution of game strategies. When we introduced the Meta-Pavlov learning to strategies as a phenotypic plasticity, population evolved to be cooperative and stable through two steps of the Baldwin effect, and was always occupied by the strategy with a modest amount of plasticity at last.

This model could be extended in a lot of directions. One obvious direction would be to attempt to apply the automatic mechanism of blending learning with evolution in the field of multi-agent systems.

References

[1] Baldwin J. M. , "A New Factor in Evolution, " *American Naturalist*, Vol. 30, pp. 441–451, 1896.

[2] Hinton G. E. and Nowlan S. J. , "How Learning Can Guide Evolution, " *Complex Systems*, Vol. 1, pp. 495–502, 1987.

[3] Anderson R. W. , "Learning and Evolution: A Quantitative Genetics Approach, " *Journal of Theoretical Biology*, Vol. 175, pp. 89–101, 1995.

[4] Sasaki T. , Tokoro M. , "Adaptation toward Changing Environments: Why Darwinian in Nature, " *Proceedings of 4th European Conference on Artificial Life*, pp. 145–153, 1997.

[5] Turney P. , Whitley D. and Anderson R.W. , "Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect, " *Evolutionary Computation*, Vol. 4, No.3, pp. 4–8, 1996.

[6] Lindgren K. , "Evolutionary Phenomena in Simple Dynamics, " *Artificial Life II*, pp. 295–311 Addison-Wesley, 1991.

[7] Nowak M. A. and Sigmund K. , "A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game," *Nature*, Vol. 364, No. 1, pp. 56–58, 1993.