

Reinforcement Learning in Auction Experiments

Shu-Heng Chen

AI-ECON Research Center
Department of Economics
National Chengchi University
Taipei, Taiwan 116
E-mail: chchen@nccu.edu.tw

Yi-Lin Hsieh

AI-ECON Research Center
Department of Economics
National Chengchi University
Taipei, Taiwan 116
E-mail: littleyam1982@yahoo.com.tw

Abstract

In this paper, we study the learning behavior possibly emerging in six series of prediction market experiments. We first find, from the experimental outcomes, that there is a general positive correlation between subjects' earning performance and their reliance on using limit order to trade. Given this connection, we, therefore, focus on subjects' learning behavior in terms of their use of limit order or market order. A 3-parameter Roth-Erev reinforcement learning (RL) model is empirically constructed for each subject. A numerical algorithm known as differential evolution is applied to estimate the maximum likelihood function derived from the RL model. The results of the estimated parameters show not just their great heterogeneity, but also the sharp contrasts among subjects. This great heterogeneity is, however, not caused by the experimental designs, such as the information distribution among agents. The statistical analysis shows that the heterogeneity comes from the subject themselves. In other words, they may be considered as personality traits of subjects. We then go further to test whether these personality traits have impact on subjects' earning performance, and some initial evidences suggest the significance of both attention and determination parameters. Hence, to be winners, not only does subjects need to learn, but also they need to learn with right personality traits (parameters).

Keyword: Reinforcement learning, Differential evolution, Prediction market experiments

1 Motivation

Learning is an multi-disciplinary area which continuously draws attention from economics, psychology, cognitive science, computer science, mathematics, and neural science. Among all proposed learning models, *reinforcement learning* seems to be one which are most commonly shared by all disciplines. In psychology, Bush and Mosteller (1955) proposed the first mathematical model of reinforcement learning, the Bush-Mosteller model has later been adapted and generalized in economics by Cross (1973), Arthur (1991), Arthur (1993), Roth and Erev (1995), and Erev and Roth (1998). Base on this development, the relevance of the reinforcement learning in economic experiments becomes available. In particular, given the fast growing of bounded-rationality literature,

reinforcement learning stands for an important class of learning. It is a kind of *reflexive learning*, i.e., learning involving little reasoning. Because of this unique position, reinforcement learning has been frequently compared with other more sophisticated learning models in either experimental economics and agent-based computational economics (Duffy, 2006).

The mathematical development of reinforcement models have constantly adding some interesting cognitive or psychological parameters to the model. These parameters, in general, are expected to be different among agents. For example, the recency effect can depends on agents' memory capacity or the associated operation mechanism of memory. However, this heterogeneity in reinforcement learning has been largely unexploited in the literature. On the contrary, since the pioneering work done by Arthur (1991) and Arthur (1993), most empirical application of reinforcement learning assumes that agents are homogeneous in these parameters. Therefore, it leaves two issues unanswered in the literature. First, if the agents' learning can be realistically represented by reinforcement learning, then how heterogeneous they can possibly be. Second, if they are heterogeneous to some extent, do those behavioral parameters matter for agents' performance. The second question just reiterates the frequently addressed question: does learning behavior matter? However, here, it does not point to different learning models, but different learning parameter values in the same model.

2 Experimental Designs

To run experiments, we used the AI-ECON Prediction Market, which was originally designed as a virtual political future market. As political future markets, agents in these experiments can trade based on their expectations of the "true value" of an asset at the expiration date, also called the *expiration price*. In the political future market, the expiration price is not known by any subjects. However, in our experiments, we introduce *insiders* to market, and release some information about the expiration price to these insiders. Motivated by Ang and Schwarz (1985), we further distinguish insiders into perfect insiders and imperfect insiders. Perfect insiders know the exact asset price at the expiration date, whereas imperfect insiders do not know the exact price, but know that a specific range covering the true price.

Totally, we have 6 different market designs. Each market experiments is composed of 20 traders. Each trader has been initially endowed with 10,000 cash and 20 shares of an asset. Each experiment comprises of 12-14 rounds. Each round lasts for 7 minutes. In each market round, traders can trade either by submitting limit orders or market orders. Short sale is not permitted. Table 1 is a summary of the experimental designs.

The six market experiments are correlated in the following way. Markets A, B and C are closely related because from A to C we increase (decrease) the number of insiders (outsiders). Markets C, D and E are closely related because from D, to C, and to E we increase (decrease) the number of perfect insiders (imperfect insiders) while keeping the number of outsiders unchanged. Markets B and F are also closely related because they are exactly same in the composition of traders, except the availability of this composition information. In all our six experiments, all traders know the existence of insiders of insiders, but, only for Market F, they do not know the exact numbers of various types of

Table 1: Experimental Designs

(1) Markets	(2) Perfect Insiders	(3) Imperfect Insiders	(4) Insiders	(5) Outsiders	(6) Market Participants	(7) Composition Information
A	0 (0%)	0 (0%)	0 (0%)	20 (100%)	20	Y
B	5 (25%)	5 (25%)	10 (50%)	10 (50%)	20	Y
C	8 (40%)	8 (40%)	16 (80%)	4 (20%)	20	Y
D	6 (30%)	10 (50%)	16 (80%)	4 (20%)	20	Y
E	10 (50%)	6 (30%)	16 (80%)	4 (20%)	20	Y
F	5 (25%)	5 (25%)	10 (50%)	10 (50%)	20	N

Insider the parenthesis gives the percentage of the respective type of traders in the market. The last column indicates whether the exact composition of trader is made known to market participants. “Y” means that the composition is a public information, and “N” means the opposite.

traders.

3 What to Learn?

In a lab with human subjects, sometimes it is difficult to be precise on what agents learned. In principle, they should have incentive to learn everything which they consider relevant to their gains or profits. However, deciding what are relevant and what are not itself is a part of learning. In this study, we assume that the main subject for agents to learn is *the use of limit order*, more exactly, the *intensity* of using limit order instead of market order. The intensity of using limited order (ILO) can be defined as in Equation (1). For each market, at the end of each round, say n , we can observe the total submission of each subject, say i . This total submission is the sum of the limit order submission (LOS) and the market order submission (MOS). The intensity of the using order limit by the i th subject on the n th round, $ILO_{i,n}$, can be defined as follows.

$$ILO_{i,n} = \frac{LOS_{i,n}}{LOS_{i,n} + MOS_{i,n}}. \quad (1)$$

Needless to say, choosing the intensity of the limit order as the learning target is certainly a simplification of the potentially more complex and multi-facet learning. Nonetheless, statistics from our experimental results (Table 2) show that the intensity of using limit order is positively correlated to the realized profits.

Let $\pi_{i,n}$ be the profit earned by subject i at the end of the n th round of one market experiment. By summing the submission and the associated profit over all rounds, one can have

$$ILO_i = \frac{\sum_{n=1}^N LOS_{i,n}}{\sum_{n=1}^N LOS_{i,n} + \sum_{n=1}^N MOS_{i,n}},$$

Table 2: Intensity of the Limit Order Submission and Profits

Market	Kendall Correlation	Spearman Correlation
A	0.568*	0.721*
B	0.720*	0.865*
C	0.717*	0.872*
D	0.727*	0.894*
E	0.765*	0.898*
F	0.377†	0.566*

The symbol * refers to the statistical significance at a significance level of 1%, and † refers to the statistical significance at a significance level of 5%.

and

$$\pi_i = \sum_{n=1}^N \pi_{i,n},$$

where N is the total number of round in that market experiment. Table 2 gives the correlation of ILO_i and π_i . To take into account the possible divergence, both the Kendall and Spearman correlations are provided, while, qualitatively, the result is not much different. They both clearly show the significant positive correlation, which indicates that individuals who used the limit order submission more intensively also tend to earn a higher profit. As a result, we have reason to believe that agents should not overlook this key.

4 The 3-Parameter RL Model: Estimation and Testing

The intuitive idea of reinforcement learning model is that for the action leading to a reward its chance of being activated again in the future will increase, whereas for the action leading to a punishment its chance will decrease. To make it operational, reinforcement learning model assigns each possible action a probability of being activated. The entire probability function over the action space is based on the *strength* of each action. The strength of an action is the its accumulated received rewards (utilities) over the past. The strength and the activation probability of each action are constantly updated by taking into account the rewards received most recently.

In literature, several different versions of reinforcement learning models have been proposed. They differ in how strength is updated and how it is mapped to the activation probability. The specific model which we consider in this paper is a version of the Roth-Erev model (Erev and Roth, 1998). In this model, the strength, $q_{i,t}$ is updated according to Equation (2).

$$q_{i,t} = \begin{cases} 1 & \text{if } t = 1, \\ (1 - \varphi) * q_{j,t-1} + (1 - \varepsilon) * \Pi_{t-1} & \text{if } i = j, t \geq 2, \\ (1 - \varphi) * q_{i,t-1} + \frac{\varepsilon}{(N-1)} * \Pi_{t-1} & \text{if } i \neq j, t \geq 2 \end{cases} \quad (2)$$

At the beginning, the strength of action is treat equally and is given by one. Then how it is updated depends on whether it has been activated. Let j be the activated action. If

$i = j$, then the strength will be updated by simply adding the payoff received from the activation, i.e., Π_{t-1} . However, two parameters, φ and ε are introduced here to make this update not that straightforward. The parameter φ is memory decaying rate. It lies between 0 and 1. If $\varphi = 1$, the strength of last period $q_{i,t-1}$ will be completely ignored (forgotten), and the strength-updating only depends on the most received payoffs. On the other hand, if $\varphi = 0$, the past strength will not decay at all, and the entire of it will roll into next period. Because of this, we call this behavioral parameter the *recency effect*.

The second parameter ε is the reserve of the own payoff. The part which is not reserved is then equally shared by other non-activated actions. ε also lies between 0 and 1. When it is 1, there is no reserve; the current received payoff is completely shared to others, while when it is 0, there is no sharing. This reserving or sharing property can then interpreted as an *attention effect*. The lower the ε , the higher the attention is put on the activated reward.

If $i \neq j$, then action i was not activated last time, and has no direct payoffs from the action. Its strength this period is then the simply the sum of the past strength which has not decayed and the payoffs received from the activated action as shown in the last part of Equation (2).

The payoff which we use in this paper is to take a logarithmic transformation of the raw payoff, as shown in Equation (3).

$$\Pi_t = \begin{cases} \ln(\pi_t) + 1 & \text{if } \pi(t) > 0, \\ 0 & \text{if } \pi(t) = 0, \\ -\ln|\pi_t| - 1 & \text{if } \pi(t) < 0 \end{cases} \quad (3)$$

The raw payoff is defined by Equation (4).

$$\pi_t = m_t + p_t \times v_t - (m_0 + p_t \times v_0) \quad (4)$$

$$p_{i,t} = \frac{\exp(\lambda * q_{i,t})}{\sum_i \exp(\lambda * q_{i,t})} \quad (5)$$

4.1 Behavioral of the Three Parameters

The three parameters to be estimated are recency effect (φ), experimental parameter (ε) and intensity of choice (λ). The parameter space for the three are $\varphi \in [0, 1]$, $\varepsilon \in [0, 1]$ and $\lambda \in [0, \infty)$. However, based on what we have discussed above, there are specific ranges for ε and λ to expect. Roughly speaking, we expect a low value of ε , and a high value of λ . Therefore, the interesting hypotheses to test are: $\mathcal{H}_0 : \varepsilon = 0$, $\mathcal{H}_0 : \varepsilon = \frac{2}{3}$, and $\mathcal{H}_0 : \lambda \gg 0$

4.2 MLE and Differential Evolution

The three parameters are estimated by the maximum likelihood estimator. The likelihood function can be written as follows.

$$\mathbf{L}(\varphi, \varepsilon, \lambda) = \prod_t p_{i,t}, \quad (6)$$

where $p_{i,t}$ is described in Equations (2) and (5). Due to the nature of $p_{i,t}$, the likelihood function is very difficult to be written in a closed form, and not to mention the analytical optimization of it. We, therefore, take a numerical algorithm known as *differential evolution* to solve the optimization problem.

4.3 Testing Zero-Intelligence

Intuitively, any minimum degree of learning should prevent our subjects from behaving randomly. Therefore, our first hypothesis is to test whether our human subject behaves randomly. To test it, we compare the likelihood from the independent random choice model with the likelihood from the 3-parameter reinforcement learning. Concretely, we test the following nulls.

$$\begin{aligned} H_0 &: L(\hat{\varphi}, \hat{\varepsilon}, \hat{\lambda}) - \left(\frac{1}{3}\right)^n \leq 0 \\ H_1 &: L(\hat{\varphi}, \hat{\varepsilon}, \hat{\lambda}) - \left(\frac{1}{3}\right)^n > 0 \end{aligned} \quad (7)$$

or

$$\begin{aligned} H_0 &: \frac{L(\hat{\varphi}, \hat{\varepsilon}, \hat{\lambda}) - \left(\frac{1}{3}\right)^n}{\left(\frac{1}{3}\right)^n} \leq 0 \\ H_1 &: \frac{L(\hat{\varphi}, \hat{\varepsilon}, \hat{\lambda}) - \left(\frac{1}{3}\right)^n}{\left(\frac{1}{3}\right)^n} > 0 \end{aligned} \quad (8)$$

where n is the number of repeated choices the subject made, which is equivalent to the number of runs in each experiment. The result shows that the null of both Equations (7) and (8) are significantly rejected¹, which implies that the subjects' choice behavior are better described by the reinforcement learning model than the zero-intelligence (purely random) model.

4.4 Heterogeneity in Agents

Even though the zero-intelligence model is rejected as a whole, it does not mean that agents are homogeneous. To see that Figure 1 gives the distribution (histogram) of the $\hat{\varphi}$, $\hat{\varepsilon}$, and $\hat{\lambda}$ over 120 agents. A few noticeable features stands out. First, the estimates are widely distributed over the parameter space, which shows a great heterogeneity of agents. Second, the dispersed distribution also indicates the existence of the extreme behavior among agents. For example, agents with strong recency attribute ($\varphi \approx 0$) coexist with agents with strong forgetting attribute ($\varphi \approx 1$); agents with strong attention control ($\varepsilon \approx 0$) coexist with agents with very weak attention control ($\varepsilon \approx \frac{2}{3}$). Likewise, agents with strong intensity of choice ($\lambda \gg 0$) are accompanied with agents with no intensity of choice ($\lambda \approx 0$).

¹See Table 6 in Section Appendix A.

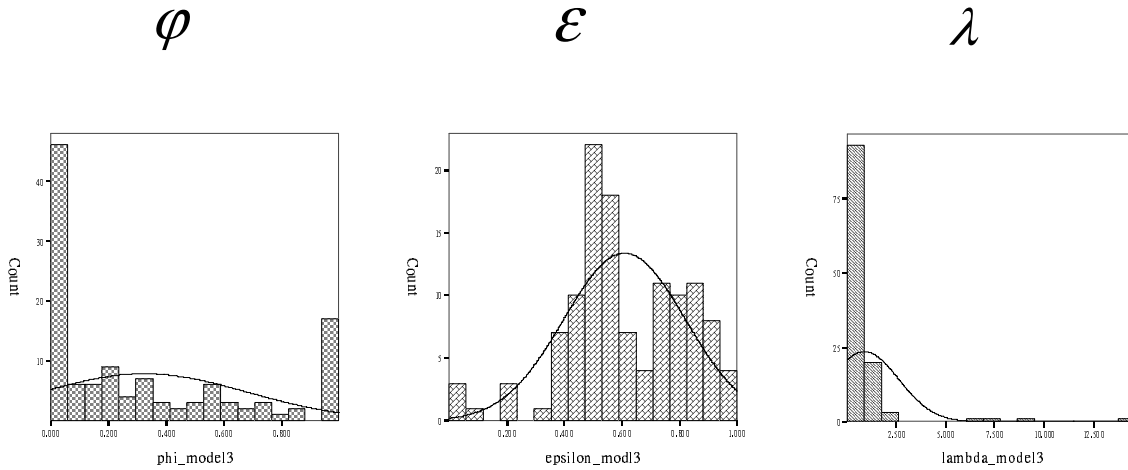


Figure 1: Distribution of the Three Estimated Parameters of All Subjects

5 Is Observed Heterogeneity Endogenous?

The picture of 120 human subjects with such great heterogeneity in their reinforcement learning behavior motivates us to think what is the cause for that. Is the heterogeneous behavior exogenous (innate) or endogenous? For the latter, we specifically ask whether different experimental settings may induce different observed learning behavior, and hence the observed agents' learning behavior may be affected by the experimental designs.

5.1 Learning Behavior under Different Markets

We first look at the learning behavior under different market experiments. Table 3 gives the mean of $\hat{\varphi}$, $\hat{\varepsilon}$, and $\hat{\lambda}$ taken over 20 agents in each market experiment. To see whether these parameters are different under different market settings, we run the Wilcoxon rank sum test for each pair of two market settings, and the results are displayed in Tables 7 to 9. The testing results consistently indicate that the observed heterogeneity in the learning parameters is not endogenously caused by different market settings.

Table 3: Means of the Three Estimated Parameters under Different Market Settings

Market	φ	ε	λ
A	0.243	0.632	0.701
B	0.320	0.560	1.211
C	0.371	0.594	1.202
D	0.287	0.650	0.567
E	0.248	0.632	0.561
F	0.473	0.581	1.058

5.2 Learning Behavior of Different Types of Traders

Even though the learning parameters are not found to be statistically different under different markets, we notice that, within each market, there are generally three different types of traders, namely, perfect insiders, imperfect insiders, and outsiders. It would then be interesting to know whether the initial information acquisition can impact the resultant learning behavior. At least, the first thought cannot exclude the possibility that the perfect insiders who know exactly where the expiration price is may have a different preference over the two trading implementations (market order and limit order) than those who are quite uncertain about it. Table 4 gives the mean of the estimated parameters with respect to the three different type of traders.

Table 4: Means of the Three Estimated Parameters under Different Types of Traders

Types	φ	ε	λ
Perfect Insiders	0.407	0.605	0.703
Imperfect Insiders	0.358	0.573	0.840
Outsiders	0.247	0.633	1.029

As before, we run the Wilcoxon rank sum test for each pair of two types of traders, and the results are shown in Tables 10 to 12. The results generally do not connect information heterogeneity to learning behavioral heterogeneity. Therefore, the observed heterogeneity in the learning behavior is neither caused by the information asymmetry. Combing this result with the previous one, we conclude that none of our two experimental settings can cause the observed heterogeneous learning behavior, which in turn rejects the endogeneity hypothesis. Therefore, subjects' differences in learning behavior, as captured by the three parameters in the fitted RL model, are likely to be exogenous (innate).

If it is subjects' personal traits which lead to the observed heterogeneity, then it is important to know whether these personal traits may actually affect their earning performance, which is shown in Table 5.

Table 5: Performance and Learning Behavior

Participants by Performance	φ	ε	λ
Top 5	0.257	0.523	1.616
Rank 6-10	0.275	0.508	0.593
Rank 11-15	0.421	0.706	0.456
Bottom 5	0.341	0.695	0.868

6 Concluding Remarks

In a process of an experiment, what the agent tried to learn and how he learned is, sometimes, not an easy issue to answer. It can become even harder in an open environment

like the prediction market, where both *the target to learn* and *the method to learn* can change over the entire learning process. However, so long as we have reason to believe that learning did happen during the experiment, then understanding learning via *learning models* remains to be a worthy research to do.

In this paper, we assume the use of the limit order as a target to learn. By further focusing on the change its intensity, a simple three-parameter reinforcement learning model is applied and estimated to fit the data. Through this simple RL model, we can easily identify some evidences of learning. First, when compared to the benchmark of zero-intelligence model, the performance of the RL model fit the data significantly better, which implies that agents did react to market feedbacks. Second, the statistical significance of the attention parameter (ϵ) being less than two thirds once confirms that agents do have a focusing attention to receiving and reacting to market feedbacks. While there are many alternative learning models to work with, a simple model like RL is good enough to make us “observe” learning and gain some insights of it.

The hypothesis that agents are heterogeneous (the heterogeneous agent paradigm), recently plays a dominating role in economic research. The simple RL model employed in this model can effectively communicate with this hypothesis. Our finding does evidence great heterogeneity among agents. Another nice virtue of RL is that it is a psychological-based model, and hence all parameters of it can be interpreted psychologically. In our case, the three parameters can be named as recency, attention, and determination. Given their psychological nature, agents’ heterogeneity can reflect their heterogeneity in personal traits, which are totally exogenous to our experimental settings. Our test does lend support to the exogeneity hypothesis. Furthermore, our analysis also shows that two of these three personal traits actually matter for agents’ earning performance.

7 Concluding Remarks

This paper provides the first application of reinforcement learning models to the choice of market order or limit order, the two basic trading operations. The most challenge issue for us is how we pose the learning problem. Given the possibility that the market dynamics can be highly intertwined with the learning dynamics, it is not entire clear whether market order is always the best solution. Fortunately, our use of reinforcement learning model does not depends on this endogeneity concern.

Appendix A Tables of Details

Table 6:

Student’s	0.0042	0.0002
Sign	< 0.0001	< 0.0001
Signed Rank	< 0.0001	< 0.0001

Table 7: Wilcoxon Rank Sum Test of Difference in Learning Behavior: φ

	A	B	C	D	E	F
A	*					
B	0.5452	*				
C	0.3268	0.7154	*			
D	0.4766	0.9035	0.4758	*		
E	0.8816	0.4424	0.2307	0.3966	*	
F	0.1141	0.3327	0.4175	0.2608	0.0860	*

Table 8: Wilcoxon Rank Sum Test of Difference in Learning Behavior: ε

	A	B	C	D	E	F
A	*					
B	0.6386	*				
C	0.4614	0.8825	*			
D	0.8825	0.5114	0.5114	*		
E	0.7574	0.3994	0.7574	0.9678	*	
F	0.8825	0.7985	0.8403	0.6773	0.4778	*

Table 9: Wilcoxon Rank Sum Test of Difference in Learning Behavior: λ

	A	B	C	D	E	F
A	*					
B	0.8613	*				
C	0.4144	0.2800	*			
D	0.6970	0.6009	0.4614	*		
E	0.4144	0.2919	0.7371	0.5463	*	
F	0.8194	0.8194	0.1413	0.5114	0.2258	*

Table 10: Wilcoxon Rank Sum Test of Difference in Learning Behavior: φ

	Insiders-I	Insiders-II	Outsiders
Insiders-I	*		
Insiders-II	0.4775	*	
Outsiders	0.0180	0.1754	*

* Insiders I: Perfect Insiders
 Insiders II: Imperfect Insiders

Table 11: Wilcoxon Rank Sum Test of Difference in Learning Behavior: ε

	Insiders-I	Insiders-II	Outsiders
Insiders-I	*		
Insiders-II	0.5662	*	
Outsiders	0.6567	0.1705	*

* Insiders I: Perfect Insiders
 Insiders II: Imperfect Insiders

Table 12: Wilcoxon Rank Sum Test of Difference in Learning Behavior: λ

	Insiders-I	Insiders-II	Outsiders
Insiders-I	*		
Insiders-II	0.3240	*	
Outsiders	0.2942	0.7214	*

* Insiders I: Perfect Insiders
 Insiders II: Imperfect Insiders

References

- Ang J. Schwarz T. (1985). Risk aversion and information structure: An experimental study of price variability in securities markets. *The Journal of Finance*, 40:825-844.
- Arthur, W.B. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *American Economic Review Papers and Proceedings*, 81(2):353-359.
- Arthur, W.B. (1993). On Designing economic agents that behave like human agents. *Journal of Evolutionary Economics* 3(1): 1-22.
- Bush R, Mosteller F (1955). *Stochastic models for learning*. New York: Wiley.
- Cross J (1973) A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87:2, 239-266.
- Duffy J (2006) Agent-based models and human subject experiments. In: Tesfatsion L, Judd K (eds), *Handbook of computational economics: Agent-based computational economics*, Vol. 2. Elsevier, Oxford, 949-1011
- Erev, I. and A. E. Roth (1998), "Predicting How People Play Games: Reinforcement Learning in Experimental Games With Unique, Mixed Strategy Equilibria," *American Economic Review*, 88, 848-881.

Roth, A. E. and I. Erev (1995), "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, Special Issue: Nobel Symposium, 8, 164-212.