Learning Rational Expectations via Policy Gradient: 
An Economic Analysis

Lawrence He

June 2019

Abstract

This paper conducts the marginal analysis and refines the updating rules for the 
adaptive learning models (e.g., Erev & Roth 1998) based on approaches in computer 
science. We propose Policy Gradient Reinforcement Learning (PGRL) to simulate 
the equilibration process of a decentralized market economy, where utility maximizers 
use only received payoffs to coordinate actions towards rational expectations. For 
each learning experience, the adaptive rules not only reinforce the chosen action by 
the marginal gain of learning, but also adjust each foregone choice by its marginal 
opportunity cost. Consequently, players exhibit risk-seeking behavior during earlier 
exploration but are risk averse later due to diminishing marginal utility of learning. 
The effectiveness of the refined rules is demonstrated through a call market simulation 
that generates diverse and complex dynamics.

JEL Classification:

**Keywords:** Rational expectations; Game Theory; Reinforcement Learning; Policy Gradient 
Theorem; Stochastic Gradient Ascent; Call Market
The most significant fact about this (price) system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take the right action.

F. A. Hayek (1945)

1. Introduction

The concept of rational expectations is fundamental in economics. In its simplest form, expectations are rational such that $U = E[\tilde{u}]$, i.e., the actual utility received ex post coincides with what is anticipated ex ante. Equilibrium game theory focuses on the right-hand side of this relation by assuming an information environment and deriving the expectation form through deductive reasoning. In doing so, how economic agents interact with the environment needs to be modeled and high rationality of the agents is required. In this paper, we focus on the left-hand side to develop a model-free sampling theory of learning rational expectations in a static game setting where players use only actual payoffs to choose actions. When learning stabilizes in equilibrium, the average realized utility approximates the expected utility by the law of large numbers. Requiring only the axiom of selfishness, we demonstrate how a rational economic order is established through tacit coordination of individuals each of whom possesses only partial knowledge (Hayek 1945).

Why do people learn? What constrain people from reaching rational expectations in a competitive market? It is because using the market institution incurs a learning cost in the process of knowledge discovery. In game theory, belief-based models assume that players possess full knowledge about the game thus can rationally deduce their best-response strategies; rational expectations arise instantly because there is no learning cost. However, when individual players observe only their own payoffs, they face uncertainties in the game thus need to constantly interact with each other in order to discover the 'knowledge of the particular circumstances of time and place', using the price system to convey information and coordinate actions (Hayek 1945). This trial-and-error process of knowledge discovery in repeated games characterizes the nature of learning. At each round of the play, each player asks himself the same question: am I satisfied with my current status? Learning will continue until his choice is regret free with certainty. Eventually, when the marginal cost of learning becomes zero (no uncertainty), players' action plans are mutually compatible (perfect coordination), everyone receives what is anticipated, and the market converges to a fully-revealing rational-expectations equilibrium (REE, Radner 1979).

\footnote{We use knowledge discovery instead of information acquisition to emphasize the subjective, tacit and contextual nature of knowledge that is to be created and discovered through human action. Meanwhile, 'information' refers to the arrival of an exogenous state variable independent of human action.}
Above economic intuition is formulated via Policy Gradient Reinforcement Learning (PGRL), which is a simulator for the equilibration process of a competitive market that operates by spontaneous order. In doing so, we connect game-theoretic reinforcement learning \cite{ErevRoth1998} to concepts and approaches in machine learning, namely, policy gradient theorem and stochastic gradient ascent \cite{SuttonBarto2018}. Policy gradient is equivalent to marginal analysis in economics, i.e., deriving the first-order conditions (F.O.C) of expected utility with respect to the vector of choice propensities as policy parameters. When the F.O.C hold, the PGRL model converges to pure strategies that are the best-response and regret-free strategies in belief-based models. Policy gradient theorem \cite{Williams1992,Suttonetal2000} guarantees the convergence by gradual movement and frequent play of the game. This theoretical link between reinforcement learning and belief-based models motivates learning in repeated games described below.

Stochastic gradient ascent is the updating rule that specifies how players update the choice propensities over the course of the game play. Specifically, at each learning experience, each player uses his received payoffs to update the vector of choice propensities. In doing so, he reinforces the chosen action $j$ with the marginal gain of learning. In the meantime, he also adjusts each unchosen action $k \neq j$ by the marginal cost of forgoing the opportunity to explore $k$. The sum (over $k$) of marginal opportunity costs defines the marginal cost of learning, which is the exact price one needs to pay to exchange action $j$. As a result, the marginal analysis depends only on actual payoffs, bypassing the challenging task of inferring the foregone payoff for each unchosen action.\footnote{In the learning literature (e.g., \cite{CamererHo1999}), foregone payoff is computed as the hypothetical payoff had the player chosen action $k$. This may require information about other players' strategies and/or assume that the player can rationally infer the hypothetical payoff. See \cite{Hoetal2007,WuBayer2015} for the challenges and methods of estimating foregone payoffs under partial information.} Then, stochastic gradient ascent adjusts the choice probabilities, which are related to the propensities through the commonly used logit function (e.g., \cite{CamererHo1999}), at each repetition of the play. When the marginal gain equals the marginal cost (i.e., both become zero), there is no further updating and the learning process converges to pure strategies in equilibrium.

The simultaneous updating of both the actual effect and the foregone effect entails rich economic interpretations of learning, whose utility quantifies each player’s subjective preference of choices as a function of learning experiences. To illustrate, the utility of learning is not determined by any prespecified functional form; instead, it is the endogenous result of joint actions of all the players throughout their learning experiences. In particular, the Law of Diminishing Marginal Utility is not an presumption nor derived from empirical observations, but is the theoretical result of selfish players gradually elevating their status to a satisfactory level. That is, the tendency toward equilibrium coupled with gradual movement ensure that,
no matter how unstable the learning process is during the early exploratory stage, it must go through an increasingly exploitive stage that approaches to the steady state of equilibrium in the limit. In the meantime, learning gradually reduces the uncertainties in the game through knowledge discovery, such that individual plans become increasingly compatible toward fully revealing REE when there is no residual uncertainty for each player.

Above theoretical results at the individual level are demonstrated through a PGRL simulation of a call market (i.e., sealed continuous double auction) where multiple traders are endowed with differential abilities to discover random realizations of a state variable. Existing studies find that neither human subjects (Pouget 2007b) nor simple reinforcement rules (Camerer et al. 2002, Pouget 2007a) are able to find the equilibrium strategies under this complex market environment. Overall, we show that the spontaneous interaction of traders results in diverse and complex individual behavior. Specifically, informed traders’ utility curves are increasing at a diminishing rate throughout their learning experiences, i.e., a standard risk averse behavior. In contrast, uninformed traders’ utility curves experience three phases: initially downward sloping due to adverse selection, upward sloping at an increasing rate when players conduct exploratory discovery (convex shaped), then upward sloping at a diminishing rate when players take exploitive actions (concave shaped). The behavior, which exhibits early risk-seeking and later risk-averse pattern with a steeper slope of the former, is not prespecified by the prospect theory (Kahneman & Tversky 1979), but endogenously arises as the result of uninformed traders receiving increasingly cooperative signals conveyed by market prices. Furthermore, for both informed and uninformed traders, their marginal utility is more volatile earlier than later, converging to zero uncertainty at the end of the repeated game when everyone discovers his own REE strategies.

In the meantime, a market-wide economic order spontaneously emerges during the equilibration process, which aggregates subjective knowledge into market data that display ex post regularities. First, the market price is increasingly revealing as individual plans become more compatible, reaching a fully-revealing steady state toward the end of trading. At that point, all the subjective components vanish and the PGRL model coincides with the predictive theory of equilibrium. Second, informed traders earn knowledge rent over the uninformed with a regular pattern. The rent is the highest when informed traders quickly discover their dominant strategies while the uninformed still conduct their random exploration. As uninformed traders continue to learn, the rent gradually shrinks to nearly zero when there is a high degree of mutual cooperation. These results thus offer a process-oriented learning explanation for the famous Grossman & Stiglitz (1980) paradox.

The rest of the paper is organized as follows. Section 2 relates the game-theoretic learning literature to approaches in machine learning. Section 3 develops the theoretical properties of
the PGRL model along with the utility-based economic interpretations. Section 4 connects the PGRL model with prominent learning models in game theory. Section 5 demonstrates the learning results at both the individual level and the market level in a simulated call market, and explains why existing reinforcement rules fail. Section 6 concludes the paper and proposes establishing an AI lab to conduct research in learning behavior.

2. Related Literature

In game theory, reinforcement learning and belief-based learning are two mainstream learning models based on different assumptions about information environment and human rationality. Belief-based models such as fictitious play (Fudenberg & Levine 1998, Chpt. 2) typically require information about opponents’ behavior for which players form beliefs in order to choose their best-response actions. Reinforcement learning is a set of heuristic, adaptive rules that require no information about other players nor the structure of the game. The rules are said to be payoff-based or completely uncoupled (Nax et al. 2016), because they depend solely on the pattern of a player’s received payoffs. There have been both theoretical work (e.g., Hopkins 2002) and empirical models (e.g., EWA of Camerer & Ho 1999) to connect the two types of learning. The PGRL model further advances their connection through rational expectations, presenting PGRL as a simulator of the equilibration process that converges to the best-response strategies of belief-based learning.

The PGRL model formalizes the economic analysis and refines the adaptive rules for the classic reinforcement learning model of Erev & Roth (1998) and others. Despite its remarkable success in describing strategic behaviors (Roth & Erev 1995, Erev & Roth 1998) and quick convergence to Nash equilibrium in market games (Erev & Rapoport 1998), why the simple reinforcement rules work and what underlies the magic of coordination remain a mystery. In fact, the extreme simplicity of reinforcement learning has caused its abandonment by cognitive psychologists (who originally proposed it) in favor of more sophisticated methods (Ho et al. 2007). Furthermore, Smith (2003) repeatedly emphasizes a ‘surprise’ in market experiments and calls for a model of process. Our economic analysis sheds new light

---

3 “These results are remarkable... How could a model that does not posit maximizing behavior of any sort and supposes that agents choose among their strategies probabilistically do better than models which postulate maximizing behavior, and that behavior is deterministic which economists typically assume? These results are all more surprising...” (Sarin & Vahid 2001, p. 105)

4 “This professional victory (in market experiments) is hollowed by the failure of standard theory to predict the ‘surprisingly’ weak conditions under which the results obtain.” (Smith 2003, p. 468)... "What is missing are models of the process whereby agents go from their initial circumstances, and dispersed information, using the algorithms of the institution to update their status, and converge (or not) to the predicted equilibrium.” (Smith 2003, p. 475)
on these issues.

Recent studies on reinforcement learning attempt to refine the payoff rules, under the names of 'aspiration' (Bendor et al. 2001, Cho & Matsui 2005), 'regret testing' (Foster & Young 2006), 'interactive trial and error' (Young 2009), 'average testing' (Arieli & Babichenko 2012), among others. These refined rules posit important asymmetries in the way that subjects respond to different patterns of experienced payoffs, which, in turn, trigger asymmetric search rules so that subjects learn desired behavior such as mutual cooperation. Compared to these random or undirected search rules, our gradient-based rules are theoretically motivated to guide the search in the steepest direction toward learning rational expectations.

The key aspect of policy gradient is that the adaptive rules depend only on the player's own choice propensities, not on those of his opponents. Instead, his interactions with other players are represented as the environment or marketplace, from which each player takes an action and receives a payoff. What coordinates the strategic behaviors of all the players is a price system (which determines individual payoffs) jointly determined by all the actions taken in the marketplace. As such, the policy gradient method explains why simple adaptive rules work and the simulation demonstrates how the magic of coordination arises as the result of spontaneous interaction of all the players.

In doing so, we embed reinforcement learning in computer science (see Sutton & Barto 2018 for the authoritative coverage) within a game-analytic framework. Despite sharing the same root in psychology and biology, reinforcement learning in game theory and that in computer science have developed along separate paths: the former aims at describing the evolution of strategic behaviors of multiple agents in relatively simple environments, whereas the later deals with the learning of a single agent in much complex environments (e.g., robotics, computer games, etc). However, there has been a general consensus that multi-agent learning with artificial intelligence would be a fruitful combination. Along this line, computer scientists have proposed inspiring concepts like Markov games (Littman 1994), Nash Q-learning (Hu & Wellman 2003), etc. However, economists are less receptive to these concepts in game theory. "The issue here, as with other sorts of reinforcement learning models, is in deciding ‘what is reinforced’, that is, what the state variable is." (Fudenberg & Levine 2007, p. 381). In light of this, the gap between the two strands of literature is mainly caused by a lack of economic interpretations of those computer science concepts and approaches. This paper aims to bridge the gap, showing how policy gradient enables a model-free sampling method that achieves the same equilibrium results as from the game-theoretic analysis. Throughout the equilibration process, what is reinforced are choice propensities.

5A special issue of Artificial Intelligence (Volume 171, 2007) was denoted to the theme 'Foundations of Multi-Agent Learning', with 12 contributions from leading game theorists and computer scientists.
by both the marginal gain and marginal opportunity cost of learning that depend only on actual payoffs. In the baseline static game with homogeneous players, the whole environment is treated as a single state in which players interact with each other. PGRL converges to pure strategies unconditionally by the refined adaptive rules. In the Bayesian game with incomplete information, however, state variables are required to represent private information possessed by each player. PGRL still converges to pure strategies conditioning on each realized value of state variables. For each player, his rivals appear to play mixed strategies as a result of his uncertainty about the private information other players condition on. We demonstrate all these results in the simulated call market.

3. Theoretical Framework

The first subsection derives the F.O.C for rational expectations in a simultaneous-move game, along with the convergence theorem to argue for the tendency toward equilibrium in repeated play of the game. The second subsection presents the PGRL model along with its economic interpretations. The static game and the notations follow Fudenberg & Kreps (1993) and the reinforcement learning notations follow Sutton & Barto (2018). The Appendix provides the technical details.

3.1 Rational Expectations and Tendency toward Equilibrium

The players are indexed by \( i = 1, \ldots, I \), and let \( -i = 1, 2, \ldots, i-1, i+1, \ldots, I \) denote the opponent players. Let \( S^i \) be the finite set of strategies for player \( i \) with typical element \( s^i \in S^i \), and \( S = S^1 \times S^2 \times \cdots \times S^I \) be the space of strategy profiles, with typical element \( s \in S \). An element of the strategies played by the opponents is denoted by \( s^{-i} \in S^{-i} \). Let \( N^i \) be the number of strategies for player \( i \), and \( N^{-i} = N^1 \times \cdots \times N^{i-1} \times N^{i+1} \times \cdots \times N^I \) be the number of all possible strategy combinations of \( i \)’s opponents. For example, if player \( i \) has two opponents and each can play two strategies, the strategy combinations are \( \{(1,1), (1,2), (2,1), (2,2)\} \) and \( N^{-i} = 4 \). Probability distribution over strategies by player \( i \) is denoted by \( p(s^i) \in p(S^i) \), and probability distribution over the opponent strategies is denoted by \( p(s^{-i}) \in p(S^{-i}) \).

Let \( u(s^i, s^{-i}) \) be the scaler payoff when \( i \) chooses \( s \) and his rivals choose \( s^{-i} \) at round \( t \). \( u(s^i, s^{-i}) \) is thus the value at the intersect \([s^i, s^{-i}]\) of the payoff matrix for \( s^i \in S^i \) and \( s^{-i} \in S^{-i} \). In this section, we assume fixed payoffs to develop the learning theory for pure strategies unconditionally, but extend to perturbed payoffs in Section 4 to incorporate pure strategies conditioning on private information.

The player forms his expectation of payoffs by summing over the joint probability of
\((s^i, s^{-i})\).

\[
E_{s^i, s^{-i} \sim p(s^i, s^{-i})} [u(s^i, s^{-i})] = \sum_{s^i, s^{-i}} p(s^i, s^{-i})u(s^i, s^{-i})
\]  

(1)

We adopt the common assumption in the learning literature (e.g., Fudenberg & Kreps 1993, Hopkins 2002) that players randomize their strategies, such that player \(i\)’s choice of strategies is independent of that of his rivals, i.e.,

\[
p(s^i, s^{-i}) = p(s^i)p(s^{-i})
\]  

(2)

The choice probability \(p(s^i)\) is to be parameterized by a probabilistic action plan or policy \(\pi_\theta(s^i) = Pr(S^i = s^i; \theta^i)\), which is differentiable with respect to a \(N^i \times 1\) vector of parameters \(\theta^i\). Each element in \(\theta^i\) is the propensity to play the corresponding strategy, and may take any bounded value, positive (i.e., reinforcing the strategy) or negative (i.e., weakening the strategy). \(p(s^{-i}) = Pr(S^{-i} = s^{-i})\) is the rivals’ joint choice probability, which is assumed to have a fixed stationary distribution.

The player strives to maximize his expected utility by choosing the optimal set of choice propensities

\[
\max_{\theta^i} J(\theta^i) = \max_{\theta^i} \sum_{s^i} \pi_\theta(s^i) \sum_{s^{-i}} p(s^{-i})u(s^i, s^{-i})
\]  

(3)

The objective function is written compactly as

\[
J(\theta^i) = \sum_{j=1}^{N^i} \pi_\theta^{i,j} U^{i,j}
\]  

(4)

where \(\pi_\theta^{i,j}\) is the player \(i\)’s probability to choose strategy \(j\) parametrized with the propensity parameter \(\theta\); and

\[
U^{i,j} = E[u(s^{i,j}, s^{-i,l})] = \sum_{l=1}^{N^{-i}} p(s^{-i,l})u(s^{i,j}, s^{-i,l})
\]  

(5)

is the expected payoff for choosing strategy \(j\), summed over the probabilities of the opponent strategies.

Marginal utility is the gradient (i.e., first-order derivatives) of expected utility with re-
spect to the vector of choice propensities,

\[ g^i \equiv \nabla_\theta J(\theta^i) = \sum_{j=1}^{N^i} \nabla_\theta \pi^i_j U^{i,j} \]  

(6)

To derive the F.O.C, \( \pi^i_j \) (player \( i \)'s probability to choose strategy \( j \)) is related to \( \theta^i_j \) (player \( i \)'s propensity to choose strategy \( j \)) through the logit function commonly used in the learning literature, such that\[ \pi^i_j = \frac{e^{\theta^i,j}}{\sum_{j'=1}^{N^i} e^{\theta^i,j'}} \]  

(7)

In Appendix A, choice \( j \)'s policy gradient or marginal utility with respect to \( \theta^i_j \) is derived as

\[ g^{i,j} = \pi^i_j \times [(1 - \pi^i_j)U^{i,j} - \sum_{k \neq j}^N \pi^i_k U^{i,k}] \]  

\[ = \pi^i_j \times [U^{i,j} - \sum_{j'=1}^{N^i} \pi^i_{j'} U^{i,j'}] \]  

(8a)

(8b)

**Expectation Theorem:** Assume that the objective function \( J(\theta^i) \) is differentiable with respect to \( \theta^i \), the F.O.C of utility maximization holds, i.e., \( g^{i,j} = 0 \) when

1. \( \pi^i_j = 0 \), i.e., there is zero probability of choosing strategy \( j \);
2. If \( \pi^i_j \neq 0 \),
   
   (a) \( (1 - \pi^i_j)U^{i,j} = \sum_{k \neq j}^{N^i} \pi^i_k U^{i,k} \), i.e., the marginal payoff of choosing strategy \( j \) equals to the marginal foregone payoff; or
   
   (b) \( U^{i,j} = \sum_{j'=1}^{N^i} \pi^i_{j'} U^{i,j'} \equiv E_{\pi^i}[U^{i}] \), i.e., the actual payoff from choosing strategy \( j \) equals to the expected payoff under policy \( \pi^i \).

Condition (2a) is standard marginal analysis, i.e., F.O.C hold when marginal received payoff equals to marginal foregone payoff. Specifically, \( U^{i,j} \) on the left-hand side is the received payoff of choosing strategy \( j \), and \( U^{i,k} \) on the right-hand side is the foregone payoff for each unchosen strategy \( k \). Thus, \( (1 - \pi^i_j)U^{i,j} \) is the marginal actual payoff, and

\( \pi^i_j \) is widely known as the exponential softmax distribution in machine learning [Sutton & Barto 2018, eq. (13.2)]. In the game-theoretic learning literature (e.g., [Camerer & Ho 1999]), choice propensities are typically multiplied by a sensitivity parameter in the exponent of the equation. This parameter is captured by the learning rate \( \alpha \) later.
\[ \sum_{k \neq j}^{N_i} \pi^{i,k}_\theta U^{i,k} \] is the marginal foregone payoff, which is the probability-weighted payoffs for all those strategies that would have been chosen. Condition (2b) is the standard rational-expectations relation: the \textit{ex post} received payoff is what the player has anticipated \textit{ex ante}. That is, the actual play (action) coincides with the intended play (strategy).

**Corollary 1:** Assume distinct expected payoffs for all the strategy choices. When the F.O.C hold, choice \( j \) is a pure strategy such that \( \pi^{i,j}_\theta = 1 \) and \( \pi^{i,k}_\theta = 0 \) for all \( k \neq j \) and \( k \in \mathbb{R}^{N_i} \).

This implies the following properties of pure strategy \( j \):

(I) Choice \( j \) is the best response to the opponents’ strategies;

(II) Choice \( j \) is regret free.

Property (I) is the standard result of belief-based models that first form beliefs of the opponent strategies then choose the best-response strategy. Regret-free property (II) means that the player is satisfied with his choice, so would not change his strategy if given another chance. Otherwise, if the player regrets his choice thus chooses another strategy with positive probability, then the strategy is no longer pure, violating Corollary 1.

**Corollary 2:** When the F.O.C hold, player \( i \) is said to find the optimal solution \( \theta^{i*} \), such that \( J(\theta^{i*}) \geq J(\theta^i) \) for any \( \theta^i \). When all players find their own optimal solutions, the game reaches rational expectations equilibrium (REE), such that \( J(\theta^{i*}) \geq J(\theta^i) \) for all \( i \in \mathbb{R}^I \).

If the F.O.C hold, Corollary 2 directly results from the properties of belief-based learning. See Fudenberg & Levine (1998) (Chpt. 2) for rigorous analysis and discussions. Intuitively, REE arises in the normal-form game as the result of homogeneous players each of whom playing pure strategies. When player \( i \) finds his own \( \theta^{i*} \), so does everyone else. The game rests in the steady state of a local maximum.

Traditionally, theoretical work to solve for \( \theta^{i*} \) focuses on the right-hand side of \( U^{i,j} = E_{\pi^j}[U^i] \). Doing so requires knowledge about the distribution of the opponent strategies \( p(s^{-i}) \) plus the assumption that player \( i \) can rationally deduce \( U^{i,j} \) in eq. (5) for all \( j \in \mathbb{R}^{N_i} \). Equilibrium game theory assumes a distribution for \( p(s^{-i}) \) so that player \( i \) forms his beliefs about the opponent strategies, then chooses his best-response strategy. Belief-based learning such as fictitious play (e.g., Fudenberg & Levine 1998 Chpt. 2) do not assume such a distribution; instead, player \( i \) learns the empirical distribution of \( p(s^{-i}) \) through a fictitious

---

8We exclude the case of indistinct payoffs. In this case, mixed strategies may arise as player \( i \) is indifferent to those equal-payoff strategies.
repeated play of the game. In contrast, reinforcement learning focuses on the left-hand side of the rational-expectations relation and requires only received payoffs.\footnote{In ‘fictitious’ play, rational expectations arise instantly just as in equilibrium models because players are assumed to possess full knowledge. One can imagine that each player obtains his best-response strategies through fictitious play before he actually plays the game with other players. This is in contrast to reinforcement learning, where players must actually play the game to acquire relevant knowledge through learning. Therefore, it is the different assumptions about the knowledge environment that distinguishes reinforcement learning from belief-based learning.}

All learning models are based on repeated play, where players are randomly matched to play the game for many rounds. At each round of the repeated play, player $i$’s propensities to choose strategy $j$ are updated by the following updating rule:

$$
\theta_{i,j}^t = \theta_{i,j}^{t-1} + \alpha \times g_{i,j}^t
$$

(9)

where $t$ is the game play at $t^{th}$ round, $\alpha > 0$ is the step size or learning rate. $g_{i,j}^t$ is a sampled value of $g_{i,j}^t$ by choosing action $j$ at round $t$, with more details given in Appendix B.

Convergence Theorem: Suppose $\alpha \to 0$ and the game is repeatedly played to horizon $T < \infty$, then $g_{i,j}^T \to 0$ in probability such that $U_{i,j}^T \to E[U_T^j]$

The theorem says that, by choosing a small learning rate and playing the game frequently, choice propensities $\theta_{i,j}^T$ will eventually stabilize at the end of the horizon. Intuitively, by the law of large numbers, the actual play (action) and the intended play (strategy) of the learning process coincide, i.e., players have learnt rational expectations by playing pure strategies. The original proof of the Convergence Theorem has been exposed in both game theory (Börgers & Sarin 1997, Proposition 1) and computer science (Sastry et al. 1994, Theorem 3.1 & 3.2). Moreover, recent advancements in machine learning (Sutton & Barto 2018, Chapter 13) have vastly generalized the theorem to model-free representations for a wide range of applications, under the name of policy gradient theorem and stochastic gradient ascent. In Appendix B, we walk through the technical details and show how the model-free sampling method lays the technological foundation for our economic analysis of learning.

To summarize, the Expectation Theorem coupled with the Convergence Theorem formalize the notion of ‘tendency toward equilibrium’, which plays a pivotal role in the properties and interpretations of the learning dynamics.

3.2 Policy Gradient Reinforcement Learning (PGRL)

At each round $t = 1, ..., T$ of a simultaneous-move game, reinforcement players take actions independently according to their respective choice distribution. At the end of round $t$,
player $i$ observes a scaler payoff $u(a^i_t, a^{-i}_t)$ when $i$ chooses $a^i$ and his rivals choose $a^{-i}$, which represents a combination of opponent actions unobservable to player $i$. To simplify notations, let $u^{i,j}_t \equiv u(a^{i,j}_t, a^{-i}_t)$ be the scaler payoff received by player $i$ taking action $j$ at round $t$.

The static game rests upon the myopic assumption commonly adopted in the learning literature [Fudenberg & Kreps 1993, Fudenberg & Levine 1998, Hopkins 2002]. By myopic, we mean that players care about only immediate payoffs, and their current actions do not affect their own future actions nor the opponent actions. Given the game setup, Appendix B derives the PGRL model as below.

The PGRL Model

For each player $i \in \mathbb{R}^I$, given initial propensities $\theta^i_0$, the game is played repeatedly for $t = 1, \ldots, T$ until convergence,

(1) **Decision Rule:** at each round $t$, each player $i$ chooses an action probabilistically according to the latest choice probabilities

$$\pi^{i,j}_{\theta_{t-1}} = \frac{\exp(\theta^{i,j}_{t-1})}{\sum_{j' = 1}^{N_i} \exp(\theta^{i,j'}_{t-1})}$$ (10)

(2) **Updating Rule:** at the end of each round, each player receives a scaler payoff $u^{i,j}_t$, and updates his own vector of choice propensities

$$\theta^i_t = \begin{cases} 
\theta^{i,j}_{t-1} + \alpha \times (1 - \pi^{i,j}_{\theta_{t-1}}) \times u^{i,j}_t & \text{for the chosen action } j \\
\theta^{i,k}_{t-1} - \alpha \times \pi^{i,k}_{\theta_{t-1}} \times u^{i,j}_t & \text{for each unchosen action } k \neq j 
\end{cases}$$ (11)

where $\alpha > 0$ is a constant learning rate or step size

The decision rule (10) is the time-varying version of choice probabilities in eq. (7). It specifies how each player chooses an action based on his existing stock of knowledge. The updating rule (11) is stochastic gradient ascent, which adjusts choice propensities by a fraction of marginal utility at each round, with the fractional amount controlled by the learning rate $\alpha$. It specified how each player accumulates the new flow of knowledge via learning by doing. What is important is the feedback loop of choice probabilities that entails economic interpretations as below.
3.2.1 Marginal Utility of Learning

In the updating rule (11), the $N^i \times 1$ vector of marginal utility at round $t$ is defined as

$$g_i^t = \begin{cases} 
(1 - \pi_{\theta_{t-1}}^{i,j}) \times u_i^{i,j} & \text{for the chosen action } j \\
-\pi_{\theta_{t-1}}^{i,k} \times u_i^{i,j} & \text{for each unchosen action } k \neq j
\end{cases}$$  \hspace{1cm} (12)$$

$g_i^t$ encodes the ‘knowledge of the particular circumstances of time and place’ (Hayek 1945, p. 521), or the flow of discovered knowledge at learning experience $t$.

**Marginal Gain of Learning (MGL) and Marginal Cost of Learning (MCL)**

The first line of eq. (12) is the *marginal gain of learning* from choosing action $j$, i.e.,

$$MGL_t^{i,j} = (1 - \pi_{\theta_{t-1}}^{i,j}) \times u_i^{i,j}$$ \hspace{1cm} (13)$$

In the meantime, by choosing $j$, the player has given up the opportunity to discover knowledge from other actions. Hence, the second line of eq. (12) is the *marginal opportunity cost* for each foregone choice $k$. The *marginal cost of learning (MCL)* is defined as the sum of the marginal opportunity costs of all foregone choices, i.e.,

$$MCL_t^{i,j} = -\sum_{k \neq j}^{N^i} \pi_{\theta_{t-1}}^{i,k} \times u_i^{i,j}$$ \hspace{1cm} (14)$$

which is the exact mirror image of the marginal gain of learning, i.e., $MGL_t = -MCL_t$ by the constraint of the probabilities that sum to one ($1 - \pi_{\theta_{t-1}}^{i,j} = \sum_{k \neq j}^{N^i} \pi_{\theta_{t-1}}^{i,k}$). The mirror image carries the economic intuition of the subjective opportunity cost: *at the margin, the utility gain from a given choice is the exact price to pay by forgoing other choices*. That is, the marginal cost of learning represents the player’s anticipated utility loss upon sacrifice of his foregone choices. Defined as such, the marginal cost of learning meets all the six criteria for the ‘Cost in a Theory of Choice’ outlined by Buchanan (1978, p. 43), which are

1. *Borne exclusively by player $i$*;
2. *Subjective*;
3. *Forward-looking*: before an action is chosen;
4. *Never to be realized*: foregone choices are never realized;
5. *Can only be measured by player $i$*;
Given the MCL definition eq. (14), we now examine the implications of the tendency toward equilibrium. As long as the marginal gain (cost) of learning is not zero, the player is not satisfied with his current action \( j \) as there is further utility gain by increasing its choice probability until it becomes a pure strategy \( \pi_{i,j}^{\theta_{T-1}} = 1 \). When this happens, the choice probability of each unchosen action \( \pi_{i,k}^{\theta_{T-1}} \) must also be zero. The convergence properties below summarize the implications.

**Convergence Properties:**

For each player \( i \), when the marginal cost of learning \( MCL_{i,j}^T \to 0 \),

1. \( MGL_{i,j}^T \to 0 \): the marginal gain of learning converges to zero, i.e., there is nothing to pay so nothing to gain;

2. \( \pi_{i,j}^{\theta_T} \to 1 \): each player has acquired full knowledge about the game, i.e., there is no uncertainty in his choice;

3. Players have correctly coordinated individual plans to be mutually compatible, so the market converges to a REE.

In the world of zero learning cost, the tendency toward equilibrium makes all the subjective components (MGL, MCL, uncertainty) vanish in the limit. And the PGRL model coincides with the steady-state equilibrium results by the Expectation Theorem.

### 3.2.2 Cumulative Utility of Learning

Let us further examine the updating rule eq. (11), which is written compactly as

\[
\theta_t^i = \theta_{t-1}^i + \alpha \times g_t^i
\]

As learning reaches round \( t \), the \( N^i \times 1 \) vector of choice propensities \( \theta_t^i \) accrues the marginal utility \( g_t^i \) for each choice up to that round. Thus, \( \theta_t^i \) represents the cumulative utility that stores the stock of knowledge up to time \( t \). Below we summarize the following two laws of learning for the dominant strategy among all the choices.

1. **The Law of Utility:** The cumulative utility is an increasing function of learning experiences;

\[\text{Sensitivity} \times \text{Experiences} = \text{Knowledge}\]
(2) The Law of Diminishing Marginal Utility: The marginal value (cost) of knowledge discovery tends to diminish as learning experiences increase.

Above laws of learning are general tendencies but may not be monotonic as there can be violations when the uncertainty is high. To illustrate, during the initial exploratory stage, knowledge discovery by trial and error may create losses thus reduce the utility of learning. As the uncertainty gradually dissipates, the cumulative utility must be increasing as players gain more experiences; otherwise, it violates the intentionality of human action. Those players who incur consistent losses are eliminated by competition during the exploratory stage, and they will not survive to the later exploitation stage. In other words, the dominant strategy for these players is ‘no action’, so we will not observe their actions nor their impact on the market data.

In addition, the tendency toward equilibrium guarantees that the marginal value of knowledge discovery must be diminishing during the later stage of learning; but it can be increasing or decreasing earlier. An increasing marginal utility may occur during the exploratory stage when the player perceives an increased level of satisfaction through aggressive learning so is willing to pay a higher marginal cost per learning experience. However, such aggressive learning will necessarily concede to conservative learning later; otherwise, it violates the Convergence Theorem.

Parameter $\alpha$

Just like the parameters in traditional utility curves (e.g., the coefficient of relative risk aversion in the CRRA utility function), $\theta^i_t$ carries a learning rate $\alpha > 0$, which is the sole parameter in the PGRL model to be fine-tuned in simulation or estimated from experimental data. $\alpha$ describes how sensitive the player is to the discovered knowledge at each learning experience. The size of $\alpha$ depends on the unit of payoffs, which can be standardized within a certain range. In theory, $\alpha \to 0$ is required by the Convergence Theorem. A large $\alpha$ leads to inaccurate approximation of expected utility and generates unstable learning results. However, a very small $\alpha$ slows down learning and may cause the learning process to be trapped at a local maximum.$^{11}$

3.2.3 Utility Index of Learning

Examining the decision rule eq. (10), we note that $\pi_{ij}$ is a monotonic transformation of $\theta^i_{ij}$ and it normalizes the cumulative utility into a probability measure. Denote by $\pi^i_{\theta_t}$ the $N^i \times 1$

$^{11}$See Kingma & Ba (2014) for the latest trend in the choice of learning rate in machine learning.
vector of choice probabilities for player $i$ at round $t$. $\pi_{\theta_t}^i$ is thus a vector of utility index for each action, measuring the player’s degree of satisfaction between zero and one.

I. Utility Index is a Cardinal Measure of Preferences of Choice

(1) *Endogenous:* it is the result of simultaneous moves by all the players up to $t$;

(2) *Subjective:* it is player $i$’s personal preferences for ranking his choices;

(3) *Decentralized:* it is dispersed among all the players $i \in \mathbb{R}^I$;

(4) *Time Varying:* the preferences of choice are relevant only for time $t$;

(5) *Continuous:* $\pi_{\theta_t}^i$ is a smooth function of $t$.

The cardinal and continuous properties carry important implications. The cardinal measure means that $\pi_{\theta_t}^i$ can not only be used to rank different choices, but the difference in ranking is also meaningful. For example, if player $i$ assigns 60% to choice $a$ and 40% to choice $b$ at time $t$, it means that $a$ is currently 20% more valuable than $b$ in the sense that $a$ is 20% more likely to be chosen. The continuous property means that the utility index is a smooth curve of time, i.e., there is no jump. Hence, this property coupled with the ‘tendency toward equilibrium’ guarantees that the equilibration process gradually stabilizes. That is, no matter what shape of the utility index curve exhibits at the early stage, it becomes increasingly stable at the later stage and eventually flattens out. Intuitively, equilibrium serves as the gravitational force that prevents player actions from going astray in permanent disorder.

II. Utility Index is a Probabilistic Measure of Choice Risk

*Knight* (1921)’s distinction between risk and uncertainty is now made clear: risk is the measurable uncertainty, and uncertainty is the risk that cannot be measured. Hence, the transformation of $\theta_t^i$ into $\pi_{\theta_t}^i$ gives rise to the Knightian notion of risk, which assigns a probability distribution to the list of choices. It reflects the player’s current riskiness in learning by doing, relevant only for player $i$ at time $t$.

Looking forward, the player may face both *exogenous uncertainty* and *endogenous uncertainty* yet to be encountered. To illustrate, exogenous uncertainty refers to the arrival of risky information determined by the nature. It is associated with the learning cost of explicit knowledge independent of human action. Endogenous uncertainty reflects players’ sheer ignorance of the game structure in that each player is unaware of his opponents’ moves so can only update his choice probabilities $\pi_{\theta_t}^i$ on the spot. It is associated with the learning cost.
of tacit knowledge that arises from the residual uncertainty in the coordination of individual action plans. The distinction between the two types of uncertainty will be made clear in the call market game.

Above distinction between risk and uncertainty allows us to view learning as a knowledge discovery process whereby players gradually resolve residual uncertainty into measurable risk that is assigned to a probability distribution. When there is no residual uncertainty, there is no more risk as players’ choices converge to pure strategies.

3.3 The Evolutionary Process of Learning

The evolutionary process of learning is described by the dynamics of the marginal utility curves and utility index curves as a function of learning experiences $t$. In the marginal utility eq. (12), the payoff value $u_{i,j}^t$ can be positive or negative but is required to have a stationary distribution, so the dynamics is not determined by different magnitude of payoffs at different times.

Note that the utility index $\pi_{i,\theta}^t$ serves as the scale of values that informs the player not only of the relative ranking of choices but also of the probability of actions to be chosen at the next round, namely, the sampling method. As the result of sampling, the player may choose the highest-ranked action (‘exploitation’), or he may end up choosing any other action in order to explore something unexperienced before (‘exploration’). It is conceivable that the tendency to explore will gradually yield to the tendency to exploit as learning reduces the residual uncertainty over time.

That players do not always exploit the highest-ranked actions, but they also attempt to explore lowed-ranked actions, is a key feature that distinguishes from traditional utility functions that pre-determine the order of choices. This non-deterministic feature gives rise to innovative activity to be described next. However, it will become gradually deterministic and tend to vanish when player actions converge to pure strategies in the limit. The evolution of the learning process is illustrated below into four stages.

(1) Random Start

Before the game starts, all the choices of player $i$ are given an equal probability of $\pi_{i,\theta_0}^j = 1/N^i$ reflecting non-informative prior uncertainty. For any $N^i > 2$, $1/N^i$ is relatively small, so both the marginal gain $(1 - 1/N^i) \times u_{i,j}^1$ and the marginal cost of learning $-(1 - 1/N^i) \times u_{i,j}^1$ are relatively large in magnitude at the first round $t = 1$. The player chooses an action randomly and may realize positive or negative payoff simply by chance.
(2) **Exploration**

Learning proceeds slowly, meaning that the stock of knowledge stored in $\theta_i^t$ accrues slightly as controlled by a small learning rate $\alpha$, so the utility index curves also evolve slowly. During the early stage, players face high uncertainties that may arise from both risky information arrival (exogenous uncertainty) and plan coordination (endogenous uncertainty). The marginal gain (cost) of learning will experience high volatility, and its values can be both positive and negative. In order to visit all possible opportunities, players tend to conduct more exploratory search than to accept their existing highest-ranked actions. The ranking of choices in their own vector of utility index changes rapidly and frequently, causing high volatility of realized payoffs in the trial-and-error process. Such volatile and unstable pattern at the early stage of learning characterizes the nature of entrepreneurial discovery \cite{Kirzner et al. 1997}.

(3) **Exploitation**

As learning proceeds, previous successful actions are reinforced more often, causing their choice probabilities to increase higher and higher, so $1 - \pi_{ij}^{t+1} \theta_{t+1}$ becomes progressively smaller over time. As a result, the marginal gain (cost) curves will become gradually stable as players gain further learning experiences, i.e., the Law of Diminishing Marginal Utility prevails. At a certain point in time, the player will find out that a particular choice tends to dominate his other choices, i.e., a dominant strategy has emerged. After that point, the player is more likely to exploit than to explore. That is, the exploratory discovery will gradually yield to the tendency to choose the highest-ranked action as uncertainties in the game diminish. This is when the stage of exploitation emerges. To be precise, we can draw a distinction line to separate 'exploration' from 'exploitation'. This is the point where the highest-ranked action (dominant strategy) reaches 50%. Before that, there is a higher probability that the player chooses any other action for exploratory purpose; but after that point, the player is more likely to choose the dominant strategy among all his choices. We thus expect an increasing and concave-shaped utility index curve for the dominant strategy.

(4) **Equilibrium**

As described by the Convergence Properties, when the market converges to REE, all the players have correctly coordinated their action plans and have learnt to play their dominant strategies with certainty. All the subjective components completely vanish and are fully incorporated into objective market data.

\cite{Kirzner et al. 1997} describes boldness, imagination, and surprise as the spirits of entrepreneurs, who are constantly seeking profitable opportunities in a fiercely competitive environment.
4. Connections with Game-Theoretic Learning Models

The PGRL model lies in between evolutionary game theory and equilibrium game theory, showing the entire process whereby players evolve from zero rationality (assumed in the former) to full rationality (assumed in the later). Below we describe the connections of the PGRL model with three prominent learning models in game theory.

4.1 Replicator Dynamics (Börgers & Sarin 1997)

The updating rule eq. (11) is closely related to the replicator dynamics (e.g., Börgers & Sarin 1997, Hopkins 2002) in evolutionary game theory. PGRL and replicator dynamics share the same theoretical properties that actual payoff converges to expected payoff by the law of large numbers. In light of this, this paper connects replicator dynamics to learning rational expectations.

Börgers & Sarin (1997) make an important connection between the population-based biological evolution and the individual learning model, showing that learning at the individual level converges to the replicator dynamics in evolutionary game theory. This paper relies on their convergence argument (Proposition 1) to motivate the notion of tendency toward equilibrium, which is essential to the law of Diminishing Marginal Utility of decentralized learning.

4.2 Cumulative Reinforcement Learning (CRL, Erev & Roth 1998)

In the CRL model, the updating rule is $\theta_{i,j}^t = \theta_{i,j}^{t-1} + u_{i,j}^t$ for the chosen action $j$. This updating rule can be seen as a special case of the PGRL with $\alpha = 1$ but no feedback loop of choice probabilities in the updating rule. That is, the chosen action is reinforced by a full amount of the received payoff, but those foregone choices are not adjusted at all.

The CRL model of Erev & Roth (1998) is motivated by the literature in psychology, where the Law of Effect and the Power Law of Practice correspond to the Law of Utility and the Law of Diminishing Marginal Utility in the PGRL model, respectively. However, the underlying mechanisms are different: while the laws are theoretical properties of the economically motivated PGRL model, they are statistically driven in the psychology-based CRL model. Specifically, what makes the CRL learning curves initially increase then flatten out is due to the accumulation of non-negative payoffs over time $t$ in the denominator of its probit probability function, so the propensities in the numerator will have a decreasing

---

13 In computer science, replicator dynamics is known as the Learning Automata (Sastry et al. 1994). The Learning Automata has exactly the same learning process as replicator dynamics.
weight as learning proceeds. However, the CRL updating rule never adjusts the propensities of unchosen actions. This means that the choice propensity for the dominant strategy may grow to a very large numeric value without bound in order to dominate the choices of other strategies. This is in sharp contrast to the PGRL model, which simultaneously updates both the actual choice and all foregone choices by the vector of time-varying choice probabilities that always sum to one.

The unilateral updating rule of the CRL cripples the equilibration analysis and causes the problem of learning saturation. That is, the probability of an initially chosen action rapidly saturate to become the dominant strategy, causing other choices to become extinct at very early stage. To solve the saturation problem, the game-theoretic learning literature introduce some ad hoc ways to randomize the choices of actions, i.e., see Nax et al. (2016) for a summary of the classified rules. In contract, the PGRL adaptive rules are theoretically motivated to guide the search in the steepest direction toward discovering rational expectations. In doing so, the refined rules endogenize the trade-off between choosing an action and forgoing others via the feedback effect of time-varying choice probabilities: the actual payoff is weighted by $1 - \pi_{i,j}^{t-1}$ for the chosen action $j$, and weighted by $-\pi_{i,k}^{t-1}$ for each foregone choice $k \neq j$. Therefore, learning saturation in the PGRL model means that choice $j$ has already converged to a pure strategy: $\pi_{i,j}^{T-1} = 1$ and $\pi_{i,k}^{T-1} = 0$ for all $k$s.

In sum, the PGRL model does not require extra parameters or ad hoc rules to randomize the choices of actions. The refined adaptive rules ensure learning proceed in the direction toward equilibrium which makes spontaneous coordination of action plans increasingly compatible.

4.3 Experience-Weight Attraction (EWA, Camerer & Ho 1999)

EWA hybridizes reinforcement learning and fictitious play through an exogenous parameter $\delta$, resulting in a combination of observed experiences and belief-based inferences. In contrast, PGRL focuses purely on observed experiences, which are shown to converge to the same best-response strategies as belief-based approaches. Furthermore, what differentiates PGRL from EWA lies in how foregone effect is computed and choice propensities are adjusted. In EWA, a hypothetical foregone payoff needs to be inferred for each unchosen action; this is not only a challenging task, but also requires information about the opponent actions and assumes high rationality of players. In contrast, PGRL does not compute foregone payoffs for unchosen

---

14 There has been ample evidence that the simple reinforcement rules of Erev & Roth (1998) fail to converge in more complex environment such as the call market (e.g., Camerer et al. 2002, Pouget 2007a).

15 Ho et al. (2007), Wu & Bayer (2015) introduce some estimation methods when players possess only partial information for the implementation of EWA.
actions; instead, it computes the marginal opportunity cost (for each foregone choice) that depends only on actual payoffs, bypassing the stringent requirements of information and rationality. As a result, the updating rules are fundamentally different. The PGRL updating rule is theoretically motivated by marginal analysis that endogenizes choice probabilities as the weights of actual payoffs, whereas the EWA updating rule is a set of heuristics that weight both actual and foregone payoffs by parameter δ along with an indicator function. In other words, the endogenous probability weighting in PGRL substitutes the role of the exogenous parameter weighting in EWA. As a result, the PGRL adaptive rules are not only theoretically sound, but also much easier to implement than the EWA rules.

EWA has been shown to nest many learning models as special cases. However, the PGRL model cannot be explicitly represented in EWA due to its the endogenous probability weighting. Moreover, it is straightforward to extend the baseline PGRL model to introduce additional parameters as in EWA. For example, the generalized updating rule can be extended to

\[ \theta_t^i = \phi \theta_{t-1}^i + \alpha \times g_t^i \]  

where \( 0 < \phi \leq 1 \) is the decay factor that depreciates past propensities; the learning rate parameter \( \alpha \) specifies how past experiences are accumulated (i.e., average utility, cumulative utility, or in between); and \( g_t^i \) is marginal utility eq. (12).

As such, the generalized PGRL model may contain \( \phi \) and \( \alpha \) in the updating rule eq. (15), plus any required parameters for initial propensities. However, it should be cautioned that adding additional parameters may bias the estimates of expectations, such that learning may not converge to REE. Therefore, whether to use the baseline updating rule eq. (11) or the generalized updating rule eq. (15) depends on the purpose. For learning rational expectations, the baseline PGRL rules should be used. On the other hand, if the purpose is to describe and predict the behavior of human subjects, the generalized rules may produce a better fit to experimental data, given numerous evidence that human plays deviate from Nash equilibrium. Thus, an immediate empirical work is to examine how the generalized PGRL performs compared to EWA and other learning models in describing experimental data.

5. A Call Market Simulation

In this section, we conduct a call market simulation to demonstrate the properties and the evolutionary process of the PGRL model in a Bayesian game setting.
5.1 Information Asymmetry and Bayesian Game

For the static game with homogeneous players considered so far, there is no explicitly specified state variables. In other words, the normal-form game has just a single state in which players interact with each other in a commonly shared marketplace. When the marginal cost of learning becomes zero, the learning process will stabilize and converge to pure strategies unconditionally. For the static game played by heterogeneous players, however, state variables need to be introduced to represent private information possessed by different types of players.\[^{16}\]

Incomplete or asymmetric information can be easily accommodated in the PGRL model by making the adaptive rules contingent upon some exogenous state variables, whose realizations are discretized into a finite set of external states. Formally, extend the $N^i \times 1$ vector $\theta^i_t$ to a $N^i \times M^i$ matrix $\Theta^i_t$, where $M^i$ is the number of possible realizations of the state variable observed by player $i$. When the state variable realizes at a certain value, he computes $g^i_t$ and updates $\Theta^i_t$ only for the corresponding column of the matrix. Assuming a stationary distribution of the state variable, the learning model will converge to a pure strategy conditioning on each state. When this occurs, player $i$ is said to find his optimal solution $\Theta^{ix}$ in a static Bayesian game. When all the players find $\Theta^{ix}$ for $i \in \mathbb{R}$ on each state, the game is said to achieve the Bayesian Nash equilibrium (Gibbons 1992).

The call market as a multilateral auction is the classic Bayesian game with incomplete information. In analytical game theory, traders are assumed to know the distribution of their rivals’ valuations, based on which they form beliefs about the rivals’ strategies in order to derive their best responses. We show how the model-free simulation method obtains the same equilibrium results without the assumptions about the opponent behavior.

5.2 Call Market Setup and Equilibrium

The setup of the call market follows Pouget (2007a) with multiple traders seeking both common value and private value of a risky asset. There are eight risk-neutral traders including four buyers and four sellers. At each trading round, each buyer (seller) submits a bid (ask) to buy (sell) one unit of the asset at an integer price in the range of $1$ to $9$, i.e., the action space contains nine elements. The common or fundamental value of the asset is either $V = 3$ or $V = 7$ that realizes randomly with equal probability at each round. To char-

\[^{16}\]This is consistent with Fudenberg & Kreps (1993) who introduce perturbed payoffs to motivate the learning behavior of mixed strategies. In their setting, player $i$ receives his own private signal whose transition is determined by the nature, and plays a pure strategy conditioning on the observed signal each round. To his rivals, player $i$ appears to play mixed strategies that reflect other players’ ignorance about his private information.
acterize information asymmetry, we assume that nature endows four informed traders (two buyers and two sellers) with the learning ability to discover the realization of $V$, whereas the remaining four traders are uninformed without such an ability. At each trading round, if any buyer’s bid is equal to or higher than any seller’s ask, a market-clearing price is called to be the midpoint of the bid and ask, and the buyer and seller transact one share at the market price; otherwise, no transaction will take place. If there are multiple bids or asks satisfying the transaction condition, the market price is set so as to maximize the total trading volume. For example, suppose two buyers bid at $6, the other two bid at $8, and two sellers offer at $4, the other two offer at $2. The market-clearing price will be $5, i.e., the midpoint of $6 and $4, so that every trader obtains a more advantageous transaction price, and the total trading volume is maximized at four shares. If multiple traders submit orders at the same price, then traders are randomly reshuffled so that everyone is given an equal chance for transaction. In addition to the risky common value ($V$) of the asset, each trader is awarded a fixed private value of $v = 0.5$ for an executed trade, so that buyers and sellers are willing to transact on the same market price to capture the private value.

It is well known that such a call market has a fully revealing perfect Bayesian equilibrium, which corresponds to the unique competitive REE [Pouget 2007a]. Specifically, informed traders will submit limit orders at the asset’s fundamental value of $V = 3$ or $V = 7$, whichever is realized; that is, their dominant strategies are to fully reveal their private information to earn the private value of $v = 0.5$. Uninformed traders, anticipating that the informed will reveal their private information, will submit aggressive limit orders: the buyers will bid at $7$ and the sellers will ask at $3$. Such buy-high and sell-low actions seem counter-intuitive. However, rational uninformed traders should be able to reason that, when informed traders adopt their dominant strategies, they can take advantage of the fully revealing prices and free ride informed traders. On the other hand, if uninformed traders submit conservative limit orders, i.e., buy at $3$ and sell at $7$, they will give up half of the trading opportunities to capture their private value; such uninformed behavior is sub-optimal. Previous studies find that both human subjects [Pouget 2007b] and ad hoc reinforcement rules [Camerer et al. 2002, Pouget 2007a] fail to discover equilibrium in such a complex environment. We thus simulate the evolutionary process for this call market to demonstrate how informed and uninformed traders coordinate their individual plans to discover their respective REE strategies.

The simulation is run for $T = 20,000$ rounds per trial, for a total number of 100 trials. Each trader is assumed a non-informative prior uncertainty, i.e., zero initial propensities so that each action has an equal probability to be chosen at the first round. We set the
learning rate $\alpha = 0.01^{17}$ and lump every 100 rounds into one run. Below we report the PGRL simulation results over 200 runs averaged across 100 trails. The evolution results are reported for both individual behavior (learning curves of marginal utility and utility index for each type of traders) and aggregate market data (informational efficiency and allocative efficiency). In all the figures, the unit on the X-axis is one run.\footnote{t-test statistics across 100 trials and mean-difference test statistics between different types of traders are not reported to save space. The figures are more intuitive and self-explanatory.}

5.3 Evolution of Individual Behavior

5.3.1 Learning Curves for Informed Traders

Figures 1.1~1.8 depict the learning curves for informed traders conditioning on the realized fundamental value $V = 3$ or $V = 7$. The marginal utility curves on the left panel plot the marginal gain of learning (the top curves) and the marginal cost of learning (the bottom curves) lumped over 100 rounds for each run. The trumpet-shaped curves show that the marginal gain of learning by choosing the dominant strategy ($V = 3$ or $V = 7$) is the exact mirror image of the marginal cost of learning, which measures the sacrificed utility of forgoing all other choices ($V = -3$ or $V = -7$) to exchange the dominant strategy. The subjective and endogenous nature of marginal utility is readily seen: informed buyers and sellers have their own marginal utility curves conditioning on different states, and they are the interactive results of simultaneous moves of all the players without a predetermined utility function. As for the time-varying patterns, the marginal gain curves on the top are generally a decreasing function of learning experiences. The curves exhibit higher volatility at the early stage, which characterize entrepreneurial discovery when informed traders face high uncertainties as visualized by the marginal cost curves at the bottom. As learning gradually reduces the residual uncertainty, dominant strategies will necessarily emerge; during the exploitation stage, the curves exhibit lower volatility and tend to flatten out to zero toward the end of 200 runs. For each marginal utility curve, what separates the stage of exploration from that of exploitation is the dividing line, which corresponds to the point where the dominant strategy reaches 50% in each utility index curve on the right panel.

The utility index curves plot choice probabilities taken at the end of each run for all the actions. To illustrate, recall that informed traders face both exogenous uncertainty and endogenous uncertainty. Exogenous uncertainty reflects the risky arrival of exogenous state variable, which realizes randomly at $V = 3$ or $V = 7$ per round. We find that informed traders, who are endowed with the ability to observe the true fundamental value,
quickly resolve the exogenous uncertainty. It takes them less than 20 runs to find out that their dominant strategies should be placing orders at $V = 3$ or $V = 7$, whichever is realized. Then the purpose of learning hereafter is to resolve the endogenous uncertainty, which arises because informed traders do not know the structure of the game, i.e., the sheer ignorance of how many other players, their strategies and payoffs, etc. For example, when one informed buyer bids at $3$, to what extent does he anticipate other traders to cooperate with his intention, rather than to front-run him by bidding at a higher price? Such tacit knowledge in plan coordination does not exist until the trader participates trading to create such knowledge on the spot that is relevant only for himself. In doing so, the buyer relies on his realized payoff to discover the new flow of knowledge which is immediately put into his existing stock.

For all informed traders, the exploration stage ends at about the $30^{th}$ run when the probability of their dominant strategies reaches 50%. The remaining 170 runs belong to the exploitation stage, during which the utility index curves of the dominant strategies experience a fast increase to 80% before the $60^{th}$ run, then it takes additional 140 runs for the curves to reach above 95% at the end. The fact that the exploitation stage occupies a much long learning period is due to the *law of diminishing marginal utility*: as knowledge discovery gradually reduces the endogenous uncertainty, the marginal gain (cost) of learning becomes increasingly smaller.

### 5.3.2 Learning Curves for Uninformed Traders

For uninformed traders, because they are not endowed with the ability to resolve exogenous uncertainty, they must rely on the cooperation from informed traders to resolve their endogenous uncertainty in the game.

Figures 2.1 & 2.3 on the left panel show the marginal utility curves for uninformed buyers and sellers without conditioning on any state variable. We find that learning creates losses for uninformed traders within the initial 10 runs. This is due to the adverse selection risk when they trade with informed traders. However, the marginal gains of learning quickly turn positive right after the $10^{th}$ run, and the curves exhibit a generally increasing trend with high volatility till $80 \sim 90$ runs, for almost half of their total learning experiences. The increasing marginal utility suggests that uninformed traders regard learning as highly valuable so they are willing to pay an increasingly higher price to acquire new knowledge. Such risk-taking behavior (i.e., risk is defined as the probability distribution of choices in Section 3.2.3) characterizes their entrepreneurial spirits in the exploration stage, and it expedites the learning process to the emergence of dominant strategies, which occurs at the $83^{th}$ run. This corresponds to the turning point of the marginal gain curves which enter into
the stage of exploitation hereafter. The marginal utility curves monotonically decrease with lower volatility, gradually closing the gap between the marginal gain and the marginal cost of learning. The Law of Diminishing Marginal Utility takes effect throughout the exploitation stage during which uninformed traders become progressively more risk averse.

Figures 2.2 & 2.4 on the right panel depict the learning process of utility index for uninformed buyers and sellers. Both S-shaped utility index curves of the dominant strategies ($V = $7 for buyers and $V = $3 for sellers) experience three phases: 1) a slight decline within the initial 10 runs (adverse selection, downward sloping); 2) increase at a faster speed up to the 83rd run (exploratory discovery, convex shaped); 3) increase with a diminishing slope throughout (exploitation of dominant strategies, concave shaped), and eventually flatten out (REE, steady state) at the end of the repeated game. As such, the early convex- and later concave- shaped utility index curves suggest that players are risk seeking during the exploration stage and risk averse during the exploitation stage. Furthermore, the slope of the convex part is steeper than that of the concave part of the curves because the Law of Diminishing Marginal Utility tends to flatten out the later part. These results closely resemble the patterns and properties of the prospect theory (Kahneman & Tversky 1979). However, compared to the state-dependent and deterministic preferences of the prospect theory, the PGRL model emphasizes the dynamic and endogenous nature of preferences. The demonstrated preferences evolve over time as the outcome of players' gradually improved learning ability to reduce the residual uncertainty; they are not pre-determined by a utility theory in whatever the functional form.

Why do uninformed traders exhibit risk-seeking behavior whereas informed traders do not? Because they are endowed with differential abilities to discover the external environment. The privileged informed traders are already at the highest level of marginal gains when they can adversely select the uninformed at the very early stage of the game. It takes them less than 30 runs for dominant strategies to emerge. For the uninformed, however, they start at a negative level of marginal utility when facing adverse selection. Their marginal utility curves must rise to a high level through aggressive exploration before the curves slope down as the market establishes its order at the 83rd run.

To summarize, viewing the learning curves of both trader groups collectively demonstrates the diversity and complexity of the learning results at the individual level. In particular, the

---

19A formal comparison of preferences between the learning theory and the prospect theory is beyond the scope of this paper. However, a possible link is conceivable. Uninformed traders are more likely to incur losses during exploratory discovery so they appear to be risk-seeking when facing losses; and they are more likely to earn stable returns during the exploitation stage so they appear to be risk averse when facing gains.

20Indeed, informed players exhibit a monotonic risk-seeking behavior during the first run (i.e., the first 100 rounds) where their marginal gains rise from a near-zero level to the highest level.
5.3.3 Coordination between Informed and Uninformed Traders

It is evident that uninformed traders learn much slower than the informed: their S-shaped utility index curves appear much flatter. During the initial 30 runs when both informed and uninformed traders conduct exploratory discovery, there is little coordination between the two groups. During that period uninformed traders’ exploratory activities are basically random, i.e., all their choice probabilities are below 20%. Coordination starts when informed traders enter into the exploitation stage, after which their behavior exhibits an increasing degree of regularity. Between 30 ∼ 60 runs, uninformed traders realize that submitting orders at the asset’s fundamental value (which they do not observe) dominate other strategies. However, whether to submit conservative limit orders (buy at $3 and sell at $7) or aggressive limit orders (buy at $7 and sell at $3) is still intriguing, suggesting that the complex environment faced by uninformed traders greatly impedes their knowledge discovery progress. The turning point occurs at about the 60th run when they learn that submitting aggressive limit orders dominates the conservative ones. Why does it take so long? Because playing aggressive strategies can be damaging for the uninformed if their anticipated cooperation from other traders falls short. Notice that at the 60th run, informed traders’ choice probabilities of the dominant strategies are all above 80%, indicating a high degree of stability in their behavior. Therefore, it is the perceived coordination that makes uninformed traders discover their dominant strategies at the 83rd run. Hereafter, both groups’ entrepreneurial spirits slowly die down, with the uninformed diminishing much quickly. Plan coordination between the two groups becomes increasingly compatible and foreseeable toward their respective REE strategies in the limit.

5.4 Evolution of Market Data

Once all the players choose their actions at each round of the repeated game, the subjective knowledge encoded in their utilities is incorporated into the market data that are objectively observed ex post. We can therefore collect historical data for the fundamental value, transaction price, realized payoffs for both informed and uninformed traders throughout the call market game. Examining the time-series and cross-sectional variations of the market data

---

21 See Pouget (2007b) for detailed analysis and why human subjects fail to play aggressive strategies in the presence of strategic uncertainty.
reveals important properties of the learning process.

5.4.1 Informational Efficiency

Informational efficiency is measured by Mean Absolute Error (MAE), i.e., the absolute difference between fundamental value and transaction price over time. MAE thus defined reflects the dollar value of the subjective knowledge that is not in the market price. In the fully revealing REE, MAE should converge to zero, i.e., subjective knowledge and objective data perfectly reconcile. Figure 3.1 (the top curve) shows an exponentially declining trend of MAE, reaching nearly zero at about the 150th run. The effect of coordination is readily seen: at the 30th run when dominant strategies already emerge for informed traders, MAE is far from zero at $0.8, because uninformed traders are still in their early stage of exploratory discovery. The formation of informationally efficient market price requires individual plans from both groups of traders to be mutually compatible. At the 83rd run when uninformed traders also discover their dominant strategies (i.e., a socio-economic order has established), MAE further reduces to $0.2. When the choice probabilities of dominant strategies are all above 90% at the 150th run, all the traders have learnt to play their dominant strategies with high probabilities and there is a nearly perfect correspondence between the subjective knowledge and the market data.

5.4.2 Allocative Efficiency and Knowledge Rent

Allocative efficiency is shown in Figure 3.2 (the middle curve) for total gains, informed gains, and uninformed gains extracted from trading as the result of mutual learning. In the fully revealing REE, each trader is able to capture his own private value of $0.5, so both trader groups should earn $2, and the total gain should be $4 per round. The gap between the gain curves reflects the knowledge rent earned by informed traders over the uninformed. Figure 3.3 (the bottom curve) plots both the marginal rent (on the left axis) and the cumulative rent (on the right axis) throughout learning experiences. To illustrate, the marginal rent increases sharply initially and reaches the highest value of $180 somewhere between the 20 and 30 runs, when both groups conduct exploratory discovery. During this period, informed traders have already resolved exogenous uncertainty by learning their way to play dominant strategies, so that they earn the equilibrium gain of $200 per run. However, uninformed traders still conduct random search as there is no coordination yet to reduce their endogenous uncertainty. They face adverse selection and trade at a loss to the informed, losing more

---

[22] This is in contrast to the cumulative reinforcement learning results (the law of actual effect only) reported by Pouget (2007a) who finds no sign of convergence at the end of learning.
than 90% ($10/$200) of their private value. After the 30th run, as uninformed traders conduct aggressive learning through exploratory discovery, their gain curve increases at a faster rate to above $100 in the mid 80 runs; in the meantime, informed traders’ marginal rent diminishes rapidly from $180 to $80. Hereafter, the rent reduction slows down by the Law of Diminishing Marginal Utility. As the trading reaches the 150th run that corresponds to the fully revealing market price, the marginal rent diminishes to nearly zero and the gains from trading stabilize at the equilibrium level throughout the remaining period. Meanwhile, the cumulative knowledge rent earned by informed traders rises from zero to $13,000 through the whole trading period.

Above analysis provides a learning explanation for the prominent Grossman & Stiglitz (1980) paradox that argues for the impossibility of informationally efficient price. Indeed, there is no incentive for informed traders to reveal their private information when the market has evolved to REE. However, such an equilibrium view is extreme in that all traders are fully rational and learn their dominant strategies instantly at no cost. It leaves no role for exploratory discovery or the orderly coordination of action plans, nor the competition process for the progressive formation of informationally efficient prices. It is both individually and socially beneficial for informed traders to reveal their information and anticipate an orderly cooperation from the uninformed. Without the participation of the uninformed, informed traders could not have earned a total of $13,000 knowledge rent, and the society can realize only half of the total private value with half of the trading volume.

5.5 Learning Results of Cumulative Reinforcement Learning (CRL, Erev & Roth 1998)

Whether or not and how traders resolve uncertainties in the game is crucial to comparing learning models. We now present the learning results using the CRL model, with the updating rule \( \theta_{t}^{i,j} = \theta_{t-1}^{i,j} + \alpha \times u_{t}^{i,j} \) where \( \alpha = 0.01 \). Informational efficiency depicted in Figure 4.1 shows that MAE does not converge to zero. Allocative efficiency, depicted in Figure 4.2, shows that neither informed nor uninformed traders realize the potential gains; the curve of total gains declines to $170 at the end of trading, i.e., less than half of the equilibrium gain ($400) per run. After 10 runs when the realized gain of informed traders reaches $200, their gain curve continues to decline to below $100 when trading ends. When informed traders do not play equilibrium strategies, uninformed traders, in response, will also deviate from equilibrium strategies to protect themselves from endogenous uncertainty. The failed

\[\text{For this updating rule with } \alpha = 1, \text{ the learning algorithm cannot run for more than 200 rounds due to numerical errors caused by extreme saturation.}\]
coordination is demonstrated by the learning curves below.

Figures 4.3 & 4.6 show that informed traders are only able to learn half set of the equilibrium strategies: buy at $3 when $V = $3 and sell at $7 when $V = $7, with both choice probabilities above 80% at the end. However, informed traders face learning difficulty on the opposite side: buy at $7 when $V = $7 and sell at $3 when $V = $3. These choices seem to converge to mixed strategies at very early stage of trading, as shown by the learning curves in Figures 4.4 & 4.5. After 20 runs, the probability to sell at $3 remains a constant at 47%, and the probability to buy at $7 remains a constant at 37%, a far distance from the equilibrium pure strategies. The probabilities of other choices also stay at a constant, suggesting that the learning process has saturated at an early stage. Furthermore, Figures 4.7 & 4.8 show that order submissions of uninformed traders are also mixed and far away from equilibrium strategies. Specifically, uninformed buyers have a 49% probability to buy at $3, 28% probability to buy at $7, 19% probability to buy at $4, etc; uninformed sellers have a 47% probability to sell at $7, 22% probability to sell at $3, 22% probability to sell at $6, etc. For both uninformed buyers and sellers, the probabilities to submit conservative orders are higher than the probabilities to submit aggressive orders. These strategies are sub-optimal.

The seemingly mixed strategies, i.e., the straight line portion of informed traders’ learning curves in Figures 4.4 & 4.5, are surprising and demand a further explanation. We identify two factors that attribute to the learning saturation. First, the naïve updating rule increments the propensities of the chosen actions rapidly to a very large numeric value, making the probabilities approach to one within 20 runs. Second, which action will eventually saturate is uncertain. At the beginning of trading, if one action happens to be reinforced more frequently, its rapid saturation tends to dominate other choices which quickly become extinct. Thus, the straight line portion of Figure 4.4 is explained as follows. For 100 total trials, buy at $6 becomes the dominant action (with the probability of one) 45 times, buy at $7 becomes the dominant action 37 times, buy at $8 becomes the dominant action 12 times, etc. Figure 4.5 is likewise.

The CRL differs from the PGRL only in how $\theta^{i,j}_t$ is updated. Thus, the comparative results highlight the importance of simultaneously updating the propensities for both the actual effect and the foregone effect. What caused learning to saturate is that when only the actual effect is reinforced, the unchosen equilibrium strategies remain unlikely to be

---

24Pouget (2007a) uses an EWA model and find that the call market converges to REE if both the actual effect and the simulated effect are reinforced. However, there is a fundamental distinction in his treatment of the simulated effect and the foregone effect of the PGRL model. Pouget assumes that adaptive traders are able to hypothetically infer the foregone payoffs of unchosen actions. In contrast, the foregone effect relies only on actual payoffs in the PGRL model.
selected. This is consistent with Pouget (2007a) who attributes the same reason to the failure of Erev & Roth (1998)'s model to discover REE.

6. Conclusion

This paper conducts a utility-based marginal analysis for the adaptive learning theories based on the policy gradient method in computer science. We propose the PGRL model to simulate the equilibration process of how rational expectations are formed in a decentralized market economy. Learning is formulated as a knowledge discovery process whereby each player exchanges the marginal gain of the chosen action with the marginal cost of all the foregone choices. Throughout learning experiences, players gradually resolve residual uncertainties (in the coordination of individual plans of the game) into risk measures, whose probability distribution guides the choice of actions by trial and error. As a result, players’ behavior evolves from earlier risk-seeking during exploratory discovery to later risk-averse when the Law of Diminishing Marginal Utility prevails. Learning proceeds until all the players are satisfied with their choices, i.e., a fully-revealing rational-expectations equilibrium has converged, when the marginal cost of learning becomes zero (no uncertainty) for everyone.

Simulating the adaptive learning process toward equilibrium is enabled by modern approaches in artificial intelligence. This paper thus advocates establishing an AI lab which employs artificial agents to conduct simulated experiments for studying learning problems in game theory. For the AI methodology to achieve its full potentials, an immediate future work is to extend the primitive version of the static game to dynamic games. In dynamic games, each player chooses an action not only from his current list of choices but also needs to consider the time dimension, i.e., how a current choice may affect future payoffs in all states, then conducts backward induction. The extension to dynamic games poses no theoretical challenge: simply replace the scaler payoff in the PGRL model with a value function under the name of ‘Nash Q-learning’ (Hu & Wellman 2003) and ‘Markov games’ (Littman 1994). One can envision the power and potentials of learning in dynamic games from recent success stories in AI.

The AI learning methodology can be useful for both equilibrium-based and process-oriented game-theoretic models. For equilibrium game theory, learning provides an inductive simulation approach to approximate Nash equilibrium solved by deductive reasoning, which typically assumes that agents possess complete information and full rationality. The

25For example, dynamic games that integrate the policy gradient and the value function form the core engine that drives the remarkable success of AlphaGo Zero, which overwhelmingly dominates human intelligence (Silver et al. 2017).
simulation approach may have a distinct advantage in complex game settings where analytical solutions can be difficult to derive, e.g., in limit order markets of Chiarella et al. (2015). Along this line, numerous research directions to extend the current pure-strategy framework, such as mixed strategies under imperfect information, multiple equilibria, cycling behavior, correlated equilibrium, etc., are left for future work.

Furthermore, the AI learning methodology provides an analytical framework for 'a model of process' (Smith 2003) whereby individuals start from their initial circumstances, using only their own payoffs to update the preferences of choice, slowly move toward their desired status. Along this line, one can use the AI lab as a test bed to examine the performance of proposed new market rules and/or state variables, and modify their game settings and implementation features in light of the test results. Understanding which equilibrium is formed, how and why it is formed, can be of immense interest to researchers and policy makers.
Appendices

A. Policy Gradient of Logit Function

Policy gradient is

\[ g^i = \nabla_\theta J(\theta^i) = \sum_{j=1}^{N^i} \nabla_\theta \pi^i_{\theta} U^i_{j} \]  (A1)

where player \( i \)'s probability of choosing strategy \( j \) is

\[ \pi^i_{\theta} = \frac{\exp(\theta^i_{j})}{\sum_{j'=1}^{N^i} \exp(\theta^i_{j'})} \]  (A2)

For arbitrary strategy \( j \) and \( k \), let \( u = \exp(\theta^i_{j}) \), \( v = \exp(\theta^i_{k}) \) and \( \Sigma = \sum_{k=1}^{N} \exp(\theta^i_{k}) \).

For \( j = k \),

\[ \frac{\partial \pi^i_{\theta}}{\partial \theta^k} = \frac{u'\Sigma - \Sigma'u}{\Sigma^2} = \frac{u\Sigma - vu}{\Sigma^2} = \pi^i_{\theta} - \pi^i_{\theta} \pi^i_{\theta} = \pi^i_{\theta} (1 - \pi^i_{\theta}) \]

For \( j \neq k \),

\[ \frac{\partial \pi^i_{\theta}}{\partial \theta^k} = \frac{u'\Sigma - \Sigma'u}{\Sigma^2} = \frac{-vu}{\Sigma^2} = \pi^i_{\theta} \pi^i_{\theta} \]

This can be written compactly as

\[ \nabla_\theta \pi^i_{\theta} = \pi^i_{\theta} \otimes (e^i_{j} - \pi^i_{\theta}) = R^i_{j,k} \]  (A3)

where

\( \otimes \) is the outer product operator;
\( \pi^i_{\theta} \) is a \( N^i \times 1 \) vector of choice probabilities for each action;
\( e^i_{j} \) is a \( N^i \) unit vector with its \( j^{th} \) element being one and zeros elsewhere;

\( R^i_{j,k} \) is a \( N^i \times N^i \) symmetric positive semi-definite matrix with the \( j^{th} \) diagonal element being \( \pi^i_{\theta} (1 - \pi^i_{\theta}) \) and its off-diagonal element being \( -\pi^i_{\theta} \pi^i_{\theta} \). \( R^i_{j,k} \) is referred to as the replicator operator, whose link to the dynamics of reinforcement learning is exposed in Börgers & Sarin (1997), Hopkins (2002) and others.

Let \( U^i \) be a \( N^i \) vector of expected payoffs. The policy gradient is

\[ g^i = \nabla_\theta \pi^i_{\theta} \times U^i = R^i_{j,k} \times U^i \]  (A4)

32
whose $j^{th}$ element is

\[
g^{i,j} = \pi^i_j \times \langle (e^{i,j} - \pi^j_i), U^{i} \rangle \tag{A5a}
\]

\[
= \pi^i_j \times \begin{bmatrix}
-\pi^i_{1}^{j} & ... & -\pi^i_{N'}^{j} \\
U^{i,1}_j & ... & U^{i,N'}_j
\end{bmatrix}
\tag{A5b}
\]

\[
= \pi^i_j \times \left( (1 - \pi^i_j)U^{i,j} - \sum_{k \neq j}^{N'} \pi^i_k U^{i,k} \right) \tag{A5c}
\]

\[
= \pi^i_j \times \left( U^{i,j} - \sum_{j' = 1}^{N'} \pi^i_{j'} U^{i,j'} \right) \tag{A5d}
\]

where $\langle \cdot, \cdot \rangle$ is the inner product of two vectors.
B. Policy Gradient Theorem and Stochastic Gradient Ascent

At each round \( t = 1, \ldots, T \), players take actions independently according to their respective strategies. At the end of round \( t \), player \( i \) observes a scalar payoff \( u(a_i^t, a_{-i}^t) \) when \( i \) chooses \( a_i^t \) and his rivals choose \( a_{-i}^t \). A realized history of actions up to time \( t \) is denoted by \( h_t = (a_1^t, \ldots, a_{t-1}^t) \), which is a typical element of the set of all possible action combinations over time, i.e., \( h_t \in H_t \). Let \( h_i^t \) be the history of actions taken by player \( i \) up to time \( t - 1 \), and \( h_{-i}^t \) be the history of actions taken by \( i \)'s opponents up to time \( t - 1 \). A complete history of actions actually taken is denoted by \( h \in H \). Thus, \( h_i \in H_i \) and \( h_{-i} \in H_{-i} \) are complete history of actions taken by player \( i \), and by his opponents, respectively. Let \( p_i(h_t; \theta_i) \) denote player \( i \)'s assessment of probability distribution of action profiles along the history \( h_t \), with the assessment rule parametrized by a vector of propensity parameters \( \theta_i \). For a realized complete history, a reinforcement learner only observes his own actions \( h_i \).

B.1 Policy Gradient of Repeated Play

Define the cumulative payoff of a complete history for player \( i \) to be

\[
   u^i(h) = \sum_{t=0}^{T-1} u(a_i^t, a_{-i}^t) \quad (B1)
\]

The player forms his expectation of cumulative payoffs by summing over the probabilities of all repeatedly played samples,

\[
   J(h; \theta^i) = E_{p^i(h; \theta^i)}[\sum_{t=0}^{T-1} u(a_i^t, a_{-i}^t)] = \sum_h p^i(h; \theta^i) u^i(h) \quad (B2)
\]

The player’s objective is to maximize his cumulative payoffs by choosing an optimal set of model parameters

\[
   \max_{\theta^i} J(h; \theta^i) = \max_{\theta^i} \sum_h p^i(h; \theta^i) u^i(h) \quad (B3)
\]

The policy gradient of the repeated play is derived as

\[
   g^i \equiv \nabla_{\theta} J(h; \theta^i) = \sum_h \nabla_{\theta} p^i(h; \theta^i) u^i(h) = \sum_h p^i(h; \theta^i) \nabla_{\theta} \log(p^i(h; \theta^i)) u^i(h) = E_{p^i(h; \theta^i)}[\nabla_{\theta} \log(p^i(h; \theta^i)) u^i(h)] \quad (B4)
\]

The third line of the gradient equation uses the log derivative trick, i.e., \( \log p^i(h; \theta^i) = \frac{\nabla_{\theta} p^i(h; \theta^i)}{p^i(h; \theta^i)} \). This is a key aspect of policy gradient theorem. It expresses the policy gradient in
terms of expectations along the history \( h \), which can be approximated by simulating a large number of sample paths or trajectories for \( h \in H \).

Given myopic behavior of all the players, the assessment of the probability of observing a given history \( h \) follows,

\[
p^i(h; \theta^i) = \prod_{t=0}^{T-1} p^i(a_i^t, a_{-i}^t; \theta^i) = \prod_{t=0}^{T-1} \pi_\theta(a_i^t)b(a_{-i}^t)
\]  \hspace{1cm} (B5)

where

\( \pi_\theta(a_i^t) = Pr(A_i^t = a_i^t; \theta^i) \) is player \( i \)’s policy of choice probability (decision rule) to choose action \( a_i^t \) at time \( t \);

\( b(a_{-i}^t) \) or \( b_{-i}^t = Pr(A_{-i}^t = a_{-i}^t) \) is player \( i \)’s probabilistic belief (assessment rule) that the joint opponent actions realizes at \( a_{-i}^t \) at time \( t \).

At the end of the repeated play, \( a_{-i}^T \) is assumed to be stationary, i.e.,

\[
p(\bar{a}_{-i}) = \lim_{T \to \infty} Pr(a_{-i}^T = \bar{a}_{-i} | a_0, \pi_\theta)
\]  \hspace{1cm} (B6)

where \( \bar{a}_{-i} \) is a steady state of opponent actions.

The key aspect from eq. (B7b) to (B7c) is that the second term drops out, i.e., the probabilistic assessment of beliefs does not enter into the policy gradient. This important property makes it possible to approximate the gradient by sampling, enabling a model-free simulation demonstrated below.

Combining eqs. (B4) and (B7c), the policy gradient for repeated play is

\[
g^i = \nabla_\theta J(h; \theta^i) = E_{a_i^t, a_{-i}^t \sim p^i(h; \theta^i)} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log(\pi_\theta(a_i^t)) \times \sum_{t=0}^{T-1} u(a_i^t, a_{-i}^t) \right]
\]  \hspace{1cm} (B8)

whose expectations can be approximated by simulating a large number of sample paths, i.e.,

\[
g^i \approx \frac{1}{L} \sum_{h \in H} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log(\pi_\theta(a_i^t)) \times \sum_{t=0}^{T-1} u(a_i^t, a_{-i}^t) \right]
\]  \hspace{1cm} (B9)

where \( L \) is the total number of sample paths.

Let us now fix a sample path, and consider the evolution of the gradient over time along
the path. We make the gradient depend on time-varying propensities, which will be updated at each round of the play as in eq. (B11a) below (see Börgers & Sarin 1997, Sastry et al. 1994, in a similar spirit). Let $g^i_t$ be the $N^i \times 1$ gradient vector at a single round $t$, we obtain from eqs. (A5a) and (B7c)

$$g^i_t = \nabla \theta \log(\pi^i_\theta) \times u(a^i_t, a^i_{t-1}) = (e^{i,j} - \pi^i_{t-1}) \times u(a^i_t, a^i_{t-1})$$

(B10)

where $e^{i,j}$ is the unit vector with its $j^{th}$ element being one and zeros elsewhere; $\pi^i_{t-1}$ is the $N^i \times 1$ vector of choice probabilities with the latest propensities $\theta^i_{t-1}$ computed up to the last round along the history $h_t$. The $N^i \times 1$ vector $(e^{i,j} - \pi^i_{t-1})$ points to the direction in parameter space that most increases the probability of repeating action $a^{i,j}$ in the future, and the scaler payoff $u(a^i_t, a^i_{t-1})$ gives the amount of the update.

B.2 Stochastic Gradient Ascent

The low-information environment of reinforcement learning allows us to drop the dependency on opponent actions in the gradient eq. (B10). Let $u^{i,j}_t \equiv u(a^{i,j}_t, a^{i-1}_t)$ be the scaler payoff received by player $i$ taking action $j$ at round $t$. At each round $t$ of the play, the player chooses an action probabilistically according to the latest policy, and observes his payoff at the end of the round. Then, the player updates the propensity vector $\theta^i$ by the following rule,

$$\theta^i_t = \theta^i_{t-1} + \alpha \times g^i_t$$

(B11a)

where

$$g^i_t = (e^{i,j} - \pi^i_{t-1}) \times u^{i,j}_t$$

(B11b)

$$\pi^i_{t-1} = \frac{\exp(\theta^{i,j}_{t-1})}{\sum_{j'=1}^{N^i} \exp(\theta^{i,j'}_{t-1})}$$

(B11c)

where $\alpha > 0$ is a constant learning rate or step size, which is the sensitivity parameter to be fine-tuned in simulation or estimated from experimental data.

Above is in the spirit of the celebrated REINFORCE algorithm of Williams (1992), which facilitates a standard implementation of the PGRL Model presented in the text. This updating rule is known as stochastic gradient ascent in machine learning. See Sutton & Barto (2018) for the method with a wide range of applications. Stochastic means to update propensities by a fraction of the gradient at each time $t$ along the sample path. This is as opposed to collecting the gradients over the entire sample path and updating propensities once at the end (batch gradient ascent), or updating propensities by a mini-batch collection of the gradients (mini-batch gradient ascent). Gradient is the synonym of marginal utility. Ascent captures the intentionality of the players who attempt to move to the direction that most increases their utility.

The Convergence Theorem says that as $\alpha \to 0$ and $T < \infty$, $g^i_T \to 0$ in probability such that $U^{i,j}_T \to E[U^{i,j}_T]$. See Börgers & Sarin 1997, Proposition 1) and Sastry et al. 1994, Theorem 3.1 & 3.2) for the original proof, and (Sutton & Barto 2018, Chapter 13) for the generalization.
References


Figure 1: Informed Traders: Marginal Utility and Utility Index
The figures below plot the marginal utility curves (on the left) and the utility index curves (on the right) for informed traders using the PGRL model where $\alpha = 0.01$. Trading is repeated for 100 trials (independent samples), with each trial running for $T = 20,000$ rounds. For each trial, the values of marginal utility are accumulated over each run, which equals 100 rounds. The values of utility index are the probability measures at the end of each run. The vertical line in each figure corresponds to the point where the choice probability of the dominant strategy reaches 50%. All the values are then averaged across 100 trials. The unit of the X-axis is one run.
Figure 2: Uninformed Traders: Marginal Utility and Utility Index
The figures below plot the marginal utility curves (on the left) and the utility index curves (on the right) for uninformed traders using the PGRL model where $\alpha = 0.01$. Trading is repeated for 100 trials (independent samples), with each trial running for $T = 20,000$ rounds. For each trial, the values of marginal utility are accumulated over each run, which equals 100 rounds. The values of utility index are the probability measures at the end of each run. The vertical line in each figure corresponds to the point where the choice probability of the dominant strategy reaches 50%. All the values are then averaged across 100 trials. The unit of the X-axis is one run.
Figure 3: Informational Efficiency, Allocative Efficiency, and Knowledge Rent

The figures below plot the curve for Mean Absolute Error (MAE) between transaction price and fundamental value (top), the curve for Gains Extracted from Trading (middle), and the curve for Knowledge Rent (bottom) using the PGRL model where $\alpha = 0.01$. Trading is repeated for 100 trials (independent samples), with each trial running for $T = 20,000$ rounds. For each trial, the MAE values are averaged and the gains and knowledge rent are accumulated over each run, which equals 100 rounds. All the values are then averaged across 100 trials. The unit of the X-axis is one run.
Figure 4: Call Market Simulation based on Classic Reinforcement Rules

The figures below plot the learning results for the reinforcement learning model of [Erev & Roth (1998)] with the updating rule $\theta_{i,j}^t = \theta_{i,j}^{t-1} + \alpha \times u_{i,j}^t$, where $\alpha = 0.01$ and $u_{i,j}^t$ is player $i$’s actual payoff by choosing action $j$. Trading is repeated for 100 trials (independent samples), with each trial running for $T = 20,000$ rounds. For each trial, the MAE values are averaged and the gain values are accumulated over each run, which equals 100 rounds. Choice probabilities are taken at the end of each run. All the values are then averaged across 100 trials. The unit of the X-axis is one run.