Unstable Dynamics of Intrinsically Motivated Learning

Arkady Zgonnikov (arkady.zgonnikov@gmail.com)
University of Aizu, Tsuruga, Ikki-machi, Aizu-wakamatsu, 965-8580 Fukushima, Japan

Ihor Lubashevsky (i-lubash@u-aizu.ac.jp)
University of Aizu, Tsuruga, Ikki-machi, Aizu-wakamatsu, 965-8580 Fukushima, Japan

Abstract

Employing the dynamical systems framework, we study the effects of intrinsic motivation on the dynamics of the learning processes. The intrinsic motivation here is the one's desire to learn not because it may cause some benefits in future, but due to the inherent joy obtained by the very process of learning. We study a simple example of a single agent adapting to unknown environment; the agent is biased by the desire to select the actions she has little information about. We show that intrinsic motivation may cause the instability of the learning process that is stable in the case of rational agent. Therefore, we suggest that the effects of human intrinsic motivation in particular and the irrationality in general may be of exceptional importance in complex sociopsychological systems and deserve much attention in the formal models of such systems.

Keywords: Mathematical modeling; decision making; learning; dynamical systems.

Introduction

Mathematical models of learning play great role in a diverse range of fields, with eminent applications found in cognitive science (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Burke, Tobler, Baddeley, & Schultz, 2010; Ahn et al., 2012), artificial intelligence (Sutton & Barto, 1998) and game theory (Fudenberg & Levine, 1998). The latter traditionally concentrates on the analysis of the Nash equilibria in games played by perfectly rational agents, thereby imposing "heroic assumptions about the knowledge and calculating abilities of the players"(Macy & Flache, 2002). It is the learning approach to game theory that addresses this issue by focusing on the adaptive behavior of the players. First, it assumes that the agents initially know little about the game context and should gradually explore the game while it is repeated indefinitely many times. Second, players base their actions solely on the previous observations; they learn by trial and error while their ultimate goal is to maximize the cumulative payoff throughout the game.

In the game learning setting it turns out that the players often fail to eventually come up with a certain efficient strategy (either pure or mixed), so their behavior can not be characterized in terms of Nash equilibria. Therefore, the inherent dynamics of learning becomes vital. A growing number of studies develop the theory behind the applications of dynamical systems to learning. Coupled replicator equations were proposed as a framework for describing the adaptive behavior of multiple learning agents interacting via a simple game (Sato, Akiyama, & Farmer, 2002; Sato & Crutchfield, 2003; Sato, Akiyama, & Crutchfield, 2005). Based on this formalism a whole range of agent behavior properties have been modelled, including noisy perception of opponent's strategies (Galla, 2009, 2011) and scale-free memory (Lubashevsky & Kanemoto, 2010). Virtually all mentioned studies emphasize that the learning process dynamics in game theoretic setting is naturally rich and non-trivial. Even the simplest systems of two agents learning to play rock-paper-scissors game may produce quasiperiodic tori, limit cycles and deterministic chaos (Sato et al., 2002; Sato & Crutchfield, 2003); the latter is often reported to be a common behavior of dynamical systems describing learning processes (Sato et al., 2005; Lubashevsky & Kanemoto, 2010; Galla & Farmer, 2013).

Indeed, the perfect rationality axiom appears unsuitable in a whole class of problems. As one may see, this fact motivated much current research on the development of the learning approach to game theory and corresponding mathematical models of learning. The canonical game theory implies that a player has full information about both the game played and the opponents faced. In contrast, the learning paradigm hypothesizes that most of this information is concealed from the players, who only possess the complete knowledge about the set of available actions and gradually learn the consequences of these actions. Even so, in the vast majority of situations studied within the learning framework so far the agents are practically assumed to be strictly rational. In other words, even learning agents still act selfishly and optimally; their rationality is bounded only in the sense of having less *a priori* information. Put within the constraints imposed by the learning paradigm, agents now have to learn the appropriate behavior strategy, but their final goal remains ultimately rational — to maximize the total payoff throughout the whole process. It means that in the course of learning the agent behavior is driven only by external factors — the actions of other players and the corresponding payoffs observed previously. In the modern dynamical models of learning the agents basically lack any kind of personality, they posess no emotions, desires or personal preferences. Up to now it is completely unknown how the dynamics of the learning would change if the agents are endowed with any kind of individuality. In the present study we face this problem.

One of the most important aspects of learning processes is the intrinsic motivation, which is commonly defined as an inspiration to do something "because it is inherently interesting or enjoyable" (Ryan & Deci, 2000). In contrast, extrinsic motivation refers to doing something "because it leads to a separable outcome" (Ryan & Deci, 2000). In relation to learning, an intrinsically motivated person learns something not (or not only) because it will lead her to a tangible reward

or payoff, but for the sake of joy obtained by the learning itself. Such person innately likes the very process of gaining new knowledge. The concept of intrinsic motivation is widely studied in psychology (Deci & Ryan, 1985; Ryan & Deci, 2000) and has vital applications in education, as well as in organizational psychology and psychotherapy. Besides, intrinsically motivated reinforcement learning (Oudeyer, Kaplan, & Hafner, 2007; Oudeyer et al., 2007; Singh, Lewis, Barto, & Sorg, 2010) is a hot topic in computer science: machine learning algorithms inspired by human cognitive processes demonstrate improved performance in a wide class of tasks. Still, despite the solid theoretical basis of intrinsically motivated learning, the dynamics of such learning processes remains a murky subject. What is the impact of the intrinsic motivation on the outcome of a learning process? Can we expect that intrinsic (and in a certain sense irrational) desire to learn will change the agent behavior substantially? Do the intrinsic motives deserve as close attention as the extrinsic ones?

Employing the dynamical systems framework, we propose a toy model capturing the effects of intrinsic motivation to learn. We study the example of a single agent facing an unknown environment, who is forced to make a repeated choice between two rewarded alternatives. The purpose of the agent is to maximize the total sum of the rewards gained throughout the process; the agent therefore should learn which of the alternatives is better rewarded. The key point of the present study is that the agent is biased: along with collecting the rewards, she also satisfies the internal need to acquire new knowledge. Therefore, the agent behavior is governed by two factors: objective (to gain as much reward as possible) and subjective (to satisfy the internal desire to learn). Our global aim in the present paper is to demonstrate on this simple example how such subjective factors may greatly impact on the dynamics of systems describing human behavior.

Model

We construct the continuous-time reinforcement learning model of a single agent adaptation, or learning, under the effect of intrinsic motivation. The discrete-time learning models is the more conventional way to describing the learning processes. However, for purposes of analysis of system dynamics the continuous models are more appropriate. We refrain from discussing the connection between the discrete-time and continuous-time reinforcement learning formulation, which is covered in detail in the literature (Sato et al., 2005; Lubashevsky & Kanemoto, 2010). We only note that the continuous-time process is actually the limit case of the discrete-time learning, when the learning agent repeatedly makes a choice infinitely many times.

In our model the agent interacts with the unknown environment by repeatedly choosing one of the two available actions $x_i,\ i \in 1,2$ and receiving corresponding reward r_i. After each decision, only the action that was actually chosen is being reinforced. In game theory it corresponds to the situation where the agent is not provided with any information about the foregone payoffs (also known as choice reinforcement (Ho, Camerer, & Chong, 2007)), in contrast to the conventional weighted fictitious play scheme. The agent accumulates the memories of the obtained rewards, and in such manner builds up an inner myopic model of the outer world. Each time the agent makes a choice she relies on the currently collected information about the quality of both actions, and, at the same time, is affected by her intrinsic motivation to learn, or to obtain new information. We interpret the latter in a sense that the agent inherently likes to select the options that add much new information to her inner model of the world. Therefore, at each instant t there are three values associated with each option x_i:

1. p_i — the probability of choosing x_i at time t
2. q_i — the agent memories about the rewards obtained in the past for selecting x_i (objective quality of x_i)
3. n_i — the novelty of the option x_i (subjective quality of x_i)

In order to complete the model, we, first, define how the choice probability p_i depends on q_i and n_i. Second, we write the equations describing time evolution of the agent memories about x_i and corresponding values of novelty.

The Boltzmann distribution (sometimes referred to as "softmax" model) fits much experimental data and is commonly used as a model for randomized human choice. We adopt it as a probability of choosing action x_i at time t

$$p_i(t) = \frac{e^{\beta[q_i+n_i]}}{\sum\limits_j e^{\beta[q_j+n_j]}}, \tag{1}$$

where $q_i + n_i$ represents the total quality of option x_i. Here without loss of generality we assume that objective and subjective factors are equally important for the agent. The constant parameter β defines to what extent the agent choice is randomized ($\beta = 0$ corresponds to the completely random choice, while $\beta = \infty$ makes the agent always select the option with the highest total quality).

We describe the evolution of the objective values $q_i,\ \ i = 1,2$ over time by the following differential equations:

$$\dot{q}_i = W(q_i,q) r_i p_i - \frac{q_i}{T_q}, \tag{2}$$

where p_i is defined by expression (1), r_i is the reward associated with action x_i. Term $r_i p_i$ can be regarded as a basic reinforcement, which is subjected to saturation effect. The term $\frac{q_i}{T_q}$ stands for the effect of the bounded capacity of the agent's memory. The events in the past separated from the present by the time considerably exceeding T_q practically do not affect the agent's behavior.

The saturation factor $W(q_i,q)$ is a weighting function depending on $q = (q_1,q_2)$. We chose $W(q_i,q)$ in such way that it bounds the infinite growth of the objective value function.

In other words, it implements the saturation effect: we tend to underestimate frequent events and overestimate rare ones. We define $W(q_i,q)$ as logistic function

$$W(q_i,q) = \frac{1}{1+e^{\frac{q_i-\overline{q}}{\gamma}}}, \quad (3)$$

where γ is the saturation coefficient and $\overline{q} = \frac{q_1+q_2}{2}$. If the current objective value of x_i is relatively large ($\frac{q_i-\overline{q}}{\gamma} \gg 1$), the probability p_i is very high and x_i is selected frequently, so the agent underestimates the reward gained: $W(q_i,q) \approx 0$. On the opposite, for the rarely selected actions p_i is low (so $\frac{q_i-\overline{q}}{\gamma} \ll 1$), and when such actions are chosen the agent pays full attention to their reward: $W(q_i,q) \approx 1$.

In order to take into account the effect of intrinsic motivation to select the options that brings much information to the agent environment model, we augment system (1–3) with the equations describing time evolution of the novelty values for each option

$$\dot{n}_i = \phi(1-p_i) - \frac{n_i}{T_n}. \quad (4)$$

Here ϕ is the parameter indicating agent's novelty rate that is the same for all of the alternative choices. In analogy to the equation (2) we define the memory capacity coefficient T_n accounting for the characteristic duration of the novelty effect.

Equations (1–4) form the basic model of the agent adaptation under the assumptions stated above. In the rest of the paper we present the preliminary analysis of the results of the numerical experiments aimed at elucidation of the basic properties of the developed model.

Numerical simulation

Prior to discussion of the numerical results, we have to underline that the similar system describing the behavior of rational agent have been analyzed previously (Sato et al., 2005). It has been elucidated that the system dynamics in case of rational agent is very simple. Namely, the agent tends to one of the equilibria depending on the system parameters, and the selected equilibria is stable with respect to the perturbations of intial conditions. Therefore, there are very few studies investigating the single agent adaptation problems, due to the absence of any complications of system behavior. We show that introducing intrinsic motivation makes the situation completely different.

Under the assumption of equal rewards ($r_1 = r_2 = 1$) we numerically simulated the dynamics of system (1–4). We discovered that depending on the values of the system parameters the structure of the system phase space trajectory may take one of two general forms: either the stable equilibrium exists or the system is unstable and has the limit cycle. We have not aimed at analyctically deriving the explicit conditions of the system instability, but the empirical observations indicate that the stable behavior is rather common, while unstable dynamics was found only for relatively narrow sets of parameters.

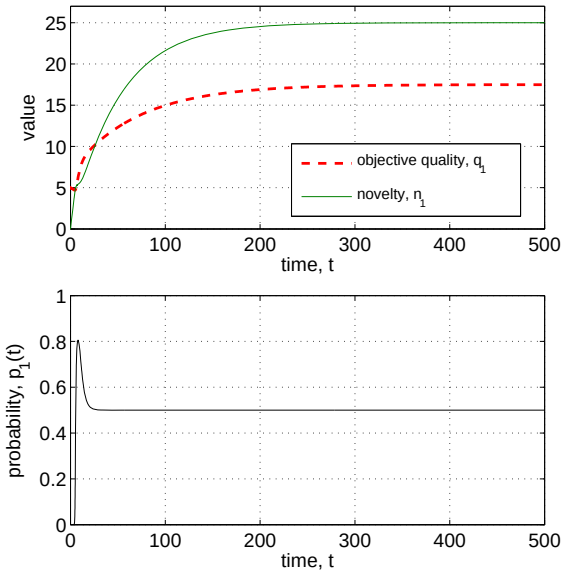

Figure 1: Stable dynamics of the analyzed system. Top frame illustrates the time evolution of the objective quality q_1 and novelty value n_1; the bottom frame represents the choice probability p_1 evolution. The time series were obtained for the time span of 500 units and following values of system parameters: $r_1 = r_2 = 1$, $\beta = 5$, $\phi = 1, \gamma = 1, T_q = 70, T_n = 50$; the initial conditions were chosen randomly.

The typical example of the stable dynamics is illustrated in Fig. 1. The agent eventually learns the mixed strategy $p_1 = p_2 = 0.5$, which is the stable equilibrium of the system. However, it is instability that often characterizes the human behavior, so we focus our attention on the second case. Fig. 2 represents the periodic motion of the system at hand. As can be seen from the top two frames, the system trajectory forms a limit cycle. Starting from the randomly selected intial values, system variables undergo periodic oscillations after a short transition process. The observed dynamical patterns correspond to the case when the decision maker changes her preferences from time to time, or, in other words, periodically "switches" from one alternative to another. The implicit dependence between the objective quality and the novelty of the option can bee seen in the bottom left frame of Fig. 2. When the quality of the alternative (as represented in the agent memories) attains local maximum, the corresponding choice probability also peaks. So this alternative is chosen frequently during some period of time and, thus, its novelty takes the lowest possible value. On the other hand, when the probability of x_i being chosen is low, the agent has relatively little information about the consequences of this action (because the memories about it eventually vanish if not reinforced regularly). Therefore, the agent intrinsic desire to learn motivates her to choose this option due to the relatively

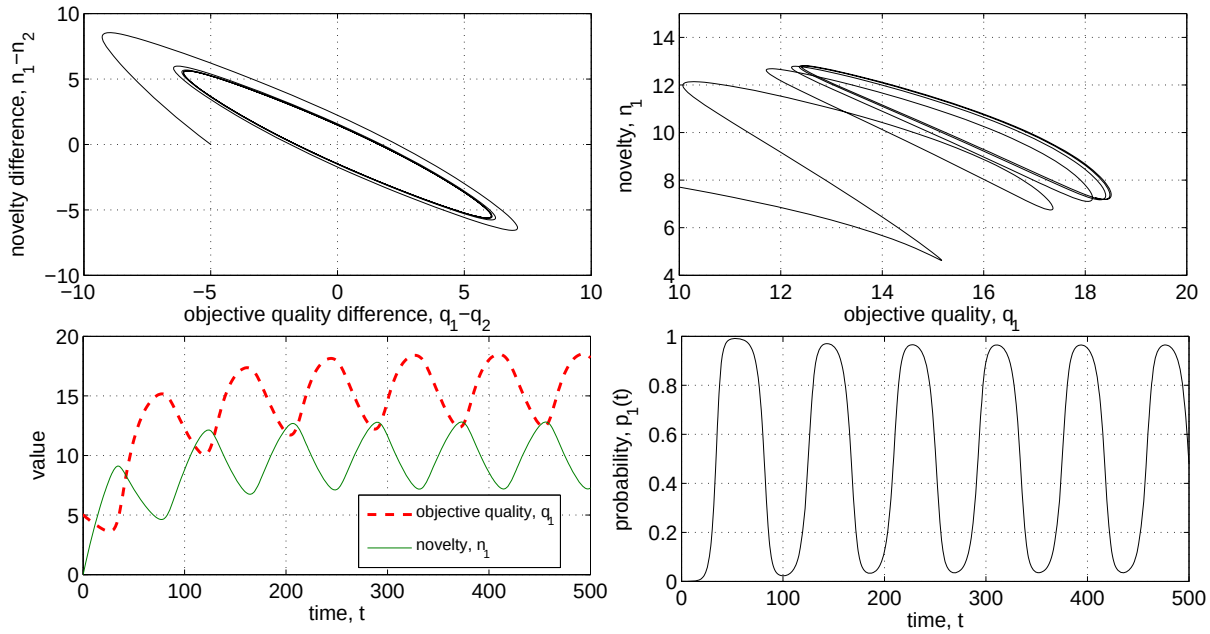

Figure 2: Unstable dynamics of the analyzed system. Top left frame illustrates the system trajectory on the plane (q_1-q_2, n_1-n_2), while top right depicts the projection of the system phase space trajectory onto the (q_1, n_1) plane. Two bottom frames demonstrate the time evolution of the objective quality q_1, novelty value n_1 (both on the bottom left frame) and the choice probability p_1 (bottom right) for option x_1. Represented results were obtained for the time span of 500 units and following values of system parameters: $r_1 = r_2 = 1,\ \beta = 2,\ \phi = 0.4, \gamma = 4, T_q = 70, T_n = 50$; the initial conditions were chosen randomly.

large amount of information that the agent might acquire.

Finally, the evolution of the choice probability $p_1(t)$ (see bottom right frame in Fig. 2) demonstrates that during considerable periods of time the probability of choosing x_1 remains close to zero; these intervals slightly precede the periods when q_1 is low and n_1 is high. Then, after staying within the vicinity of zero, the probability rapidly reaches the maximum value around unity and in turn remains near this value for the next half-cycle.

The conducted numerical analysis confirms that the the system (1–4) actually exhibits the properties one may intuitively anticipate from the intrinsically motivated agent. The agent learns one of the optimal options, but being biased she eventually tends to discard the established strategy that proved its efficiency in favor of the novel one. Moreover, the preliminary analysis of the non-symmetric case revealed that the similar behavior can be observed even when the rewards are not equal. This fact requires a thorough investigation and will be reported elsewhere.

The results presented in the present work already enable us to conclude that even the simplest systems with boundedly rational agents may exhibit non-trivial dynamics. However, more detailed analysis of the proposed model is required. Particularly, the system stability conditions are still to be determined. Also under the scope of future work is the question of how the system dynamics patterns depend on the sysem parameters, namely, the novelty rate, perception thresholds and the parameters characterizing the capacity of the agent memory.

Conclusion

We have proposed a dynamical model of intrinsically motivated learning. In the various learning models developed previously in game theory and cognitive science the learning subject is assumed to act rationally in achieving the ultimate goal — to maximize the cumulative reward gained during the learning. We challenge this approach by assigning a piece of non-rationality to the learning agent. The curiosity is what biases the selfish agent in our model.

We confine our scope to the case of single agent adaptation and follow the reinforcement learning setting. The agent behavior in our model is governed by two stimuli. The objective stimulus is traditional — to maximize the total payoff collected throughout the process. The subjective one is irrational — to engage in active learning as much as possible, because the very learning process is enjoyable. We show that the agent biased in such way at least under some conditions does not stick to the optimal strategy of behavior, in contrast to the rational learning agent. Rather, in such cases the agent preference continuously varies in an oscillatory way. Performing the simple numerical analysis of the model, we demonstrate that the intrinsic motivation leads to the instability of the learning dynamics.

Our results give evidence to the fact that the intrinsic mo-

tivation in particular and the bounded rationality in general may cause the significant changes in the behavior of single- and multi-agent systems. We argue that the intrinsic motives should be paid no less attention than the extrinsic ones, if one considers the systems where human decisions are of the primary importance.

References

Ahn, W.-Y., Rass, O., Shin, Y.-W., Busemeyer, J., Brown, J., & O'Donnell, B. (2012). Emotion-based reinforcement learning. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 124–129).

Burke, C., Tobler, P., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, *107*(32), 14431–14436.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer.

Fudenberg, D., & Levine, D. (1998). *The theory of learning in games*. MIT press.

Galla, T. (2009). Intrinsic noise in game dynamical learning. *Physical review letters*, *103*(19), 198702.

Galla, T. (2011). Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment*, *2011*(08), P08007.

Galla, T., & Farmer, J. (2013). Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, *110*(4), 1232–1236.

Ho, T., Camerer, C., & Chong, J. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, *133*(1), 177–198.

Lubashevsky, I., & Kanemoto, S. (2010). Scale-free memory model for multiagent reinforcement learning. Mean field approximation and rock-paper-scissors dynamics. *The European Physical Journal B-Condensed Matter and Complex Systems*, *76*(1), 69–85.

Macy, M., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, *99*(Suppl 3), 7229–7236.

Oudeyer, P., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation, IEEE Transactions on*, *11*(2), 265–286.

Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54–67.

Sato, Y., Akiyama, E., & Crutchfield, J. (2005). Stability and diversity in collective adaptation. *Physica D: Nonlinear Phenomena*, *210*(1), 21–57.

Sato, Y., Akiyama, E., & Farmer, J. (2002). Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, *99*(7), 4748–4751.

Sato, Y., & Crutchfield, J. (2003). Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, *67*(1), 015206.

Singh, S., Lewis, R., Barto, A., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *Autonomous Mental Development, IEEE Transactions on*, *2*(2), 70–82.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). Cambridge Univ Press.