Neural Networks 62 (2015) 83–90

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2015 Special Issue

Immediate return preference emerged from a synaptic learning rule for return maximization

Yoshiya Yamaguchi [a], Takeshi Aihara [a,b], Yutaka Sakai [a,b,*]

[a] *Graduate School of Brain Sciences, Tamagawa University, Tokyo, Japan*
[b] *Brain Science Institute, Tamagawa University, Tokyo, Japan*

ARTICLE INFO

Article history:
Available online 14 May 2014

Keywords:
Inter-temporal choice
Delay discount
Reinforcement learning
Synaptic plasticity

ABSTRACT

Animals including human often prefer immediate returns to larger delayed returns. It holds true in the human communications. Standard interpretation of the immediate return preference is that an animal might subjectively discount the value of a delayed reward, and that might choose the larger valued one. The interpretation has been successfully applied to explain behavior of many species including human. However, the description is not necessarily sufficient to apply for interactions of individuals. This study adopts a different approach to seek a possibility that immediate return preference may be reproduced by learning rule to maximize objective outcomes. We show that a synaptic learning rule to achieve the temporal difference (TD) learning for outcome maximization fails the maximization and exhibits immediate return preference if the context is not properly represented as a internal state.

1. Introduction

Animals, including human, often prefer immediate returns to delayed ones (Ainslie, 1974; Hwang, Kim, & Lee, 2009; Kobayashi & Schultz, 2008; Mazur, 1987; Mazur & Biondi, 2009; Richards, Mitchell, de Wit, & Seiden, 1997). It holds true in human communications. We are apt to desire immediate reply of our partner in a conversation. In iterative games, the balance of valuation of immediate and delayed returns affects the interactive performance (Masuda & Ohtsuki, 2009; Yoshida, Dolan, & Friston, 2008). How long future returns they evaluate is a significant question to understand communication and interaction of individuals.

Preferences for delayed returns of nonhuman animals are examined in iterative inter-temporal reward choice tasks (Fig. 1(A)). Alternative responses are associated with different amounts and delays of rewards. The total duration from a response to the start of the next trial is fixed (L in Fig. 1(A)). Therefore, the response associated with the larger amount results in the larger total returns, independently of the delays. However, animals' choices depend on the pattern of the delays.

Standard interpretation of the immediate return preference is that an animal might subjectively discount the value of a delayed reward, and that might choose the larger valued one. Subjective value V estimated from animal's preference is a monotonically decreasing function of the time delay D to the reward, and it is often approximated by the hyperbolic function,

$$V \simeq \frac{R}{1+\kappa D}, \tag{1}$$

where R denotes the physical amount of reward, and parameter κ is a positive constant that represents the degree of the discounting value of delayed reward. The nature of delay discounting has been studied in the fields of behavioral economics and experimental psychology, and often discussed in relation to impulsivity or addiction (Acheson, Vincent, Sorocco, & Lovallo, 2011; Becker & Murphy, 1988; Kim & Lee, 2011; Takahashi, 2011). Possible origins (Nakahara & Kaveri, 2010; Takahashi, 2005) and biological significance (Sozou, 1998) of discounted subjective value have been discussed in the literature, but is still open to discussion.

For the consistency, the subjective value of rewards given at multiple timings, V_{total}, should support additivity,

$$V_{\text{total}} = \sum_k V(R_k, D_k), \tag{2}$$

where $V(R_k, D_k)$ is the subjective value of a reward of amount R_k and delay D_k. Suppose the additivity (2), nearly hyperbolic discount would predict rational preference in iterated trial tasks, because the integration of a hyperbolic function diverges, and hence the contributions of posterior rewards are dominant. However,

[*] Corresponding author at: Brain Science Institute, Tamagawa University, Tokyo, Japan. Tel.: +81 42 739 8400; fax: +81 42 739 8400.
E-mail address: sakai@eng.tamagawa.ac.jp (Y. Sakai).

http://dx.doi.org/10.1016/j.neunet.2014.04.004
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

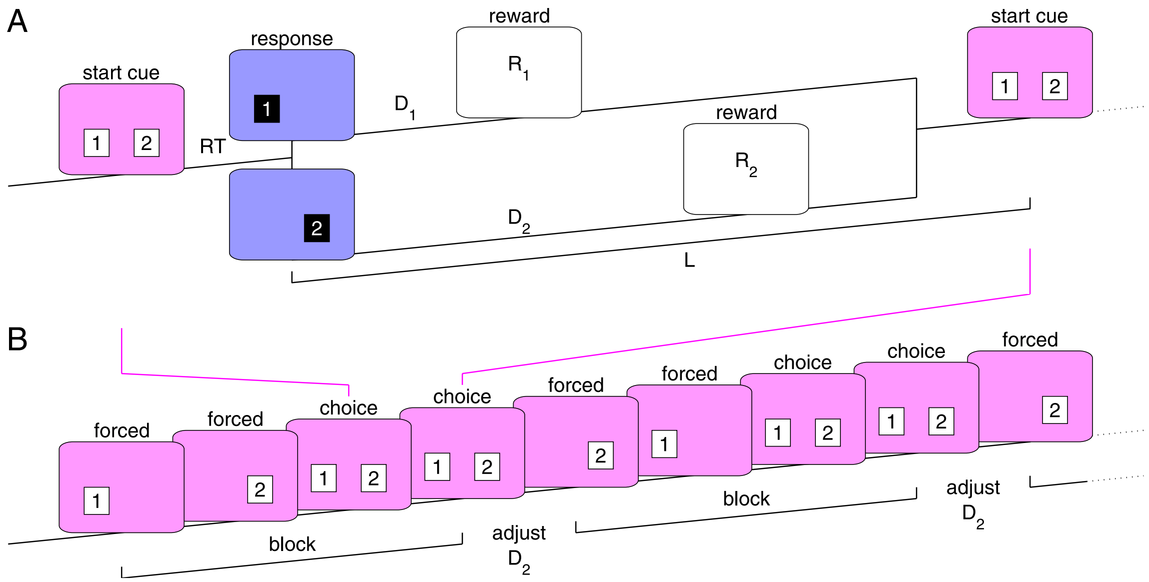

Fig. 1. Procedure of an inter-temporal choice task. (A) Time course of a trial. (B) Example trial sequence.

animals exhibit immediate return preference even in iterated trial tasks (Kobayashi & Schultz, 2008; Mazur, 1987; Mazur & Biondi, 2009). The hyperbolic discount conflicts with the additivity.

Another interpretation is that an animal might care only about the time from its response to the return, and maximize the foraging rate in the bounded span, independently of the blank after the return. The bounded foraging rate is equal to $R/(\epsilon + D)$, where ϵ denotes the consuming time of the reward. It corresponds to the case that $\kappa = 1/\epsilon$ in Eq. (1). This is simple and consistent with animals' behavior. However, the biological significance of neglect of blanks after returns is unknown.

In this paper, we point out another possibility that an animal might attempt to maximize the total returns, but fail to complete in some reason. To make a suitable response for a situation by situation, an animal should apprehend the current situation from all sensory stimuli and the recent history. Since available information is infinite, it is in general a hard problem to select information relevant to the future returns. Hence, an animal may not appropriately apprehend the current situation, especially in artificial experiments. For instance, in an inter-temporal choice task (Fig. 1), it is unknown how an animal might apprehend the situation during the delay period, and whether the context of the chosen response might be reflected on the internal state. We attempt to examine whether such failures in internal state representation can be a cause of the immediate return preference.

2. Theory

Maximization of the total returns

At every time t, a learning system is given sensory stimuli and choose one response $A(t) = a$ from the available responses $\{a = 1, 2, \ldots\}$ or no response $A(t) = 0$. To make an appropriate choice for a given situation, the system needs to apprehend the current state from available information that consists of the past history of sensory stimuli and responses. The current state is represented by internal variable $S(t)$ in the system. Here, we assume that possible states are discrete. Hence, switching of state is a point process. In a state $S(t) = s$, each response a is stochastically chosen with a state-dependent response rate ρ_{as}.

As the result of responses, the individual may get some food or pay some energy costs for responses and consumption. Let $r(t)$ denote the instantaneous return rate at time t. The amount value of food and costs given from time t' to t'' is written as the integration, $\int_{t'}^{t''} r(t)dt$. The present framework supposes that the stochastic rule of the environment is temporally invariant.

Standard frameworks of reinforcement learning presuppose the Markov property or ergodic property (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). In contrast, the present framework does not presuppose these properties because these cannot be determined only by the environment given to an animal, but do depend on the definition of state space that is determined by the animal. For given environment, the present framework assumes only the temporal invariance of the stochastic rule. For the state space determined by an animal, the present framework assumes only the discreteness. Because of the discreteness in the continuous time framework, the manner of state transitions is the same as the semi-Markov decision process (Bertsekas, 1995; Daw, Courville, & Touretzky, 2006; Sutton & Barto, 1998), which adopts the discrete time steps with variable intervals by means of real time. However, the learning agent is allowed to response at any timings in the present framework, while the agent should make a choice only at given timings in the framework of the semi-Markov decision process.

The learning system attempts to optimize the set of state-dependent response rates $\{\rho_{as}\}$ to maximize the average return rate $E[r(t)]$. A simple method for maximization is the gradient method. For a fixed set of response rates $\{\rho_{as}\}$, the average return rate is determined. The gradient for ρ_{as} is obtained (see Appendix A) as

$$\frac{\partial E[r(t)]}{\partial \rho_{as}} = \int_0^\infty d\tau \Big(E\Big[r(t+\tau) | \mathcal{C}_{as}(t) \Big] - E\Big[r(t+\tau) | S(t) = s \Big] \Big) P_s, \qquad (3)$$

where $\mathcal{C}_{as}(t)$ denotes the stochastic condition that response a occurs at time t in state $S(t) = s$, and the probability distribution of state $P_s \equiv \Pr\Big(S(t) = s\Big)$ does not depend on time t because of the temporally invariant stochastic rule.

TD learning and the limitation

Since the gradient (3) is described as a infinite integration, it is difficult to estimate through practical trial-and-error learning.