Journal of Mathematical Psychology 66 (2015) 59–69

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior

Kentaro Katahira

Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan

HIGHLIGHTS

- Reinforcement learning (RL) models and regression models have been used for choice data analysis.
- We investigated the relation between these two approaches.
- We found a special case in which an RL model is equivalent to a regression model.
- Based on the relation, we discuss how the RL parameters are related to history dependence.

ARTICLE INFO

Article history:
Received 7 November 2014
Received in revised form
19 February 2015
Available online 15 May 2015

Keywords:
Reinforcement learning
History dependence
Regression model
Model-based analysis

ABSTRACT

Reinforcement learning (RL) models have been widely used to analyze the choice behavior of humans and other animals in a broad range of fields, including psychology and neuroscience. Linear regression-based models that explicitly represent how reward and choice history influences future choices have also been used to model choice behavior. While both approaches have been used independently, the relation between the two models has not been explicitly described. The aim of the present study is to describe this relation and investigate how the parameters in the RL model mediate the effects of reward and choice history on future choices. To achieve these aims, we performed analytical calculations and numerical simulations. First, we describe a special case in which the RL and regression models can provide equivalent predictions of future choices. The general properties of the RL model are discussed as a departure from this special case. We clarify the role of the RL-model parameters, specifically, the learning rate, inverse temperature, and outcome value (also referred to as the reward value, reward sensitivity, or motivational value), in the formation of history dependence.

1. Introduction

Reinforcement learning (RL) models have been widely used to analyze the choice behavior of living systems in a wide range of behavioral studies, including psychology and neuroscience (Corrado & Doya, 2007; Daw, 2011; O'Doherty et al., 2004; O'Doherty, Hampton, & Kim, 2007; Yechiam, Busemeyer, Stout, & Bechara, 2005). Evidence of the neural correlates for the subcomponents assumed in RL theory (e.g., reward prediction error, action value) provides the validity to perform an RL model-based analysis for choice behavior (Niv, 2009; Samejima, Ueda, Doya, & Kimura, 2005; Schultz, 1997).

An essential feature of the RL model is the formulation of what action to take based on previous experiences of reward or punishment regarding the action. In addition, the linear regression-based approach, using reward history and choice history as explanatory variables and future choice as an objective variable, has also been used to model choice behavior (Corrado, Sugrue, Seung, & Newsome, 2005; Katahira, Fujimura, Okanoya, & Okada, 2011; Kovach et al., 2012; Lau & Glimcher, 2005; Seo, Barraclough, & Lee, 2009; Seo & Lee, 2009; Seymour, Daw, Roiser, Dayan, & Dolan, 2012; Sugrue, Corrado, & Newsome, 2004). The linear regression approach is useful for estimating how reward and choice histories influence future action (e.g., how much influence the reward from n trials ago has on future actions). However, the relation between the RL model and regression models has not been explicitly addressed. Specifically, to what extent and how the predictions differ between the two models has not been explored. Hence, the dependence on reward history in RL models has not been clearly described.

In the present study, we aimed to clarify the relation between the parameters of the RL model and the influence of reward history on future choice (specifically, the regression coefficients of the logistic regression models). Because the regression model can

E-mail address: katahira@lit.nagoya-u.ac.jp.

http://dx.doi.org/10.1016/j.jmp.2015.03.006

directly represent the dependency on the reward and choice histories, investigating the relationship between RL-model parameters and regression models would provide valuable information about which behavioral factors may underlie the differences in the model parameters. Conversely, using the relation, one can predict which types of the behavioral sequences can be expected given a specific set of model parameters. To achieve these aims, we performed analytical calculations and numerical simulations. We focused on fundamental RL-model parameters: the learning rate, the outcome value (also referred to as the reward value, reward sensitivity, or motivational value), and the inverse temperature (also referred to as the exploration parameter). These parameters have been used to characterize how psychological factors or personality traits of individuals affect choice behavior (Katahira, Fujimura, Matsuda, Okanoya, & Okada, 2014; Katahira et al., 2011; Kunisato et al., 2012; Lindström, Selbing, Molapour, & Olsson, 2014). However, how these parameters are related to particular behavioral aspects has not been explored sufficiently. The present study will aid in the interpretation of the different impacts of the RL-model parameters.

In the present study, we focus on probabilistic learning tasks (also called bandit problems), in which a decision-maker must choose between a set of options, each with different unknown reward rates, to maximize the total reward. The reward rates can dynamically change during the task, but they do not depend on past choices. Such probabilistic learning tasks have been widely used in psychology and neuroscience research. Simplified Q-learning models have often been used in RL model-based analysis of data obtained using this task. A general Q-learning model computes the action value, which is an expected future reward, for each "state" (Watkins & Dayan, 1992). However, for the probabilistic learning tasks that we consider here, there is only one state, and thus, a state variable is not required. Thus, in this study, we focus on a simplified Q-learning model without a state variable, and we will refer to this model as simply the "Q-learning model".

In this paper, we first introduce several variants of Q-learning models for probabilistic learning tasks. Next, we describe a logistic regression model, which is a typical regression model used to analyze choice data. Among the variants of the Q-learning models, we find that the forgetting Q-learning model (F-Q model), in which the value of an unchosen option decays by the same amount as the value of chosen, non-rewarded option, is able to make predictions equivalent to those of the logistic regression model. We can view the general Q-learning model as a model that deviates from this special case. The deviation clarifies the special properties of standard RL models. We then present numerical simulation results that demonstrate the relation between the parameters in Q-learning models and the history dependence of choice. Finally, we discuss several implications of our results.

2. Models

2.1. Reinforcement learning models

Here, we introduce an RL model (Sutton & Barto, 1998). Specifically, we consider the Q-learning model (Watkins & Dayan, 1992), which is the most commonly used model for model-based analysis of choice behavior. Throughout the paper, we consider a case with only two options; however, our results can be generalized to multiple-option cases. The model assigns each action, i, an action value, $Q_i(t)$, where t is the index of the trial. In the default setting, the initial action values, $Q_i(1)$, are set to zero, i.e., $Q_1(1) = Q_2(1) = 0$. Let $a(t) \in \{1, 2\}$ denote the option that was chosen at trial t. Based on the set of action values, the model computes the probability of choosing option 1 using the soft max function:

$$P(a(t) = 1) = \frac{\exp(\beta Q_1(t))}{\exp(\beta Q_1(t)) + \exp(\beta Q_2(t))} \tag{1}$$

$$= \frac{1}{1 + \exp(-\beta [Q_1(t) - Q_2(t)])}, \tag{2}$$

where β is the inverse temperature parameter that determines the sensitivity of the choice probabilities to difference in values. The model subsequently evaluates the outcome of the action. The outcome value in trial t is denoted by $R(t)$. We typically simply set the binary value for $R(t)$ such that $R(t) = 1$ if a reward is given and $R(t) = 0$ if no reward is given. The impact of different outcomes may be quantified by choosing parameters $R(t) = \kappa_1$ if outcome 1 is given, $R(t) = \kappa_2$ if outcome 2 is given, and $R(t) = 0$ if a control outcome is given (Katahira et al., 2014, 2011, 2015).

Based on the outcome, the action values for the chosen option i are updated as follows:

$$Q_i(t+1) = Q_i(t) + \alpha_L (R(t) - Q_i(t)), \tag{3}$$

where α_L is the learning rate that determines how much the model updates the action value depending on the reward prediction error, $R(t) - Q_i(t)$. For the unchosen option j ($i \neq j$), the action value is updated as follows:

$$Q_j(t+1) = Q_j(t) - \alpha_F Q_j(t) \tag{4}$$

$$= (1 - \alpha_F) Q_j(t), \tag{5}$$

where α_F is the forgetting rate (Ito & Doya, 2009). In a common RL model-based analysis, the action value of the unchosen option is not typically updated. This convention can be represented by setting $\alpha_F = 0$. We call this the standard Q-learning model. In this study, the forgetting rate parameter plays an important role in the identification of the connection between the regression and RL models, as discussed later.

2.2. Linear regression models

Next, we will introduce a regression model that predicts a choice from the reward and choice history of previous trials (Corrado et al., 2005; Lau & Glimcher, 2005; Sugrue et al., 2004). Here, we consider a binary outcome case such that $R(t) = 1$ when the reward is given and $R(t) = 0$ when no reward is given. Following the convention of Corrado et al. (2005) and Lau and Glimcher (2005), we represent the reward history $r(t)$ as follows:

$$r(t) = \begin{cases} 1 & \text{if option 1 is chosen and a reward is given at trial } t, \\ -1 & \text{if option 2 is chosen and a reward is given at trial } t, \\ 0 & \text{if no reward is given at trial } t. \end{cases}$$

We represent the choice history $c(t)$ as follows:

$$c(t) = \begin{cases} 1 & \text{if option 1 is chosen at trial } t, \\ -1 & \text{if option 2 is chosen at trial } t. \end{cases}$$

With these history variables, the regression model is defined with a predictor:

$$h(t) = \sum_{m=1}^{M_r} b_r(m) r(t-m) + \sum_{m=1}^{M_c} b_c(m) c(t-m), \tag{6}$$

where $b_r(m)$ and $b_c(m)$ are the regression coefficients for the trial m trials ago. The constants M_r and M_c are the history length for the reward history and the choice history (from the past trials to the current trial), respectively. Sugrue et al. (2004) and Corrado et al. (2005) used a linear regression approach with an identity-link function and optimized the regression coefficients so that they