Are Standard Reinforcement Learning Models too Flexible?

Patrick J. Rice (pjrice@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Mathi Manavalan (mathi@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Andrea Stocco (stocco@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Keywords: Reinforcement learning; Model architecture; Behavioral modeling.

Introduction

In the past two decades, neuropsychological research into the cognitive bases of learning and behavior has increasingly benefited from the application of computational models of learning, such as those derived from reinforcement learning (RL) theory. Despite advances in RL, many studies continue to rely on the older Rescorla-Wagner (RW) learning model. While the RW model is missing many of the more modern RL features, it is still applied in an attempt to describe multiple aspects of brain functioning and participant behavior such as ERP dynamics related to response and feedback (Cavanagh, Frank, Klein, & Allen,2010). Here, we demonstrate that under a simple target-discrimination/stop signal task, three RL model variants with increasing constraints are indistinguishable in terms of fit to participant data, despite converging to different regions of the parameter space.

Reinforcement Learning Models

Model Architectures

We implemented three RL models ("single-update", "double-update", and "targeted-update") to model participant behavior under a target-discrimination/stop-signal task. Participants had to learn the correct stimulus-response mappings through trial-and-error while monitoring for potential stop signals, resulting in "Go" and "Stop" trials (see Reinhart & Woodman,2014 for additional task details). Each model utilized a standard update rule:

$$Q(s_{t+1},a_{t+1}) = Q(s_t,a_t) + \alpha\delta_t \tag{1}$$

Where $Q(s_t,a_t)$ is the Q-value associated with performing action a in state s at time t, α is a parameter that controls the rate of learning, and δ_t is defined as:

$$\delta_t = [r_{t+1} - Q(s_t,a_t)] \tag{2}$$

These estimated Q-values are transformed into a distribution of probability of selection over the range of possible actions on any given trial through a softmax action selection rule:

$$P(a) = \frac{e^{Q_t(a)/\beta}}{\sum_{b=1}^{n} e^{Q_t(b)/\beta}} \tag{3}$$

These three equations comprise the entirety of the single-update model.

The double-update model is almost identical to the single-update model, with the additional assumption that reward under the task is anti-correlated. That is, if taking one action generates positive reward, then any other action would have generated negative reward (and vice-versa). This assumption allows the model to make a second update on each trial, applying the opposite of the reward ("antiReward") that was received to every action that was not taken. While uncommon, this updating approach has been utilized to some success (Reiter et al.,2016).

However, human participants generally begin with some knowledge regarding the dynamics of a new task, such as through instructions given in a lab setting. As such, we created a third model that attempted to encode two pre-existing expectations: that "Go" trials should be responded to, while "Stop" trials should not be responded to. To encode these expectations, model updates on any given trial were "targeted" so that positive/negative reward was more appropriately allocated to the response options.

Under standard initialization conditions, all three models have only two free parameters, the learning rate α and the noise in action selection β.

Model Initialization

In RL modeling, Q-values are typically initialized as "0" for every potential state-action pairing (standard initialization) so that every potential action is equally probable before any learning occurs. An alternative manner of encoding initial expectations (the goal of the "targeted" model) is to initialize some state-action pairings with a nonzero value. We took this approach by estimating a third parameter "initVal" for each of the three models, representing some negative value that two general state-action pairings are initialized at: responding to "Stop" trials, and not responding to "Go" trials (alternative initialization).

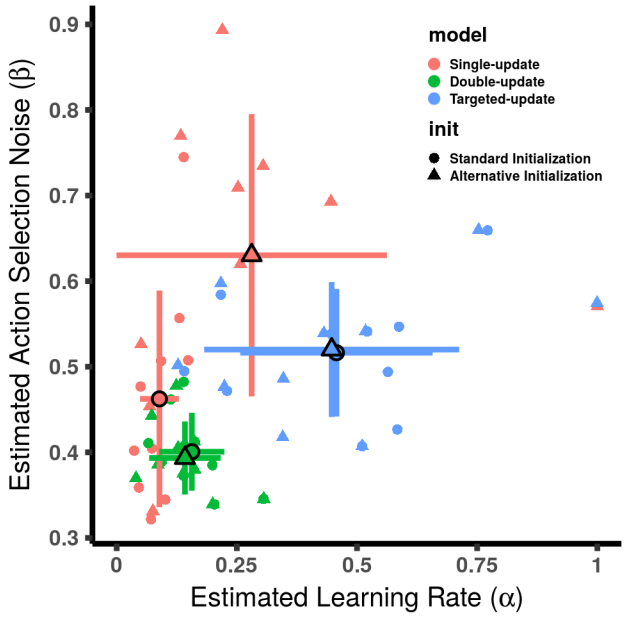

Figure 1: Estimated α versus β parameter of each participant for the three model architectures, under both standard and alternative initialization conditions. The points circumscribed in black are mean parameter estimates across participants. Horizontal and vertical lines indicate standard deviation for the α and β parameters, respectively.

Results and Discussion

Differences in model fits (pseudo-R^2) and parameter estimates were examined through Welch's paired-samples *t*-testing. We observed no differences in model fit between both model architectures (single/double/targeted updates) and initialization approaches (standard/alternative).

Comparing estimated learning rates (α parameter) between model architectures initialized in the standard manner revealed that the double-update model's α was significantly greater than that of the single-update model (paired $t(14.7) = -2.70$, $p = 0.017$), while the targeted-update model's α was significantly greater than that of the double-update model (paired $t(11) = -4.51$, $p < 0.001$). However, when comparing model architectures under the alternative initialization protocol, the double-update model's α was no different than that of the single-update model (paired $t(10.3) = 1.5$, $p = 0.16$), while the targeted-update model's α was again significantly greater than that of the double-update model (paired $t(10.4) = -3.5$, $p = 0.005$). When comparing between model initialization protocols, no significant differences in estimated α was found. This suggests that our alternate initialization procedure conveys information to the model that it would quickly learn through double-updating; when both are present, no additional benefit is gained. However, comparing the learning rate α of the "targeted-update" model to that of the "double-update" model makes clear that the "targeted" nature of the updates speeds learning above and beyond that of alternative initialization/double-updating.

Under standard initialization, the estimated noise in action selection (β parameter) was significantly greater for the targeted-update model, when compared to the double-update model (paired $t(14.9) = -4.2$, $p < 0.001$). For alternative initialization, the single-update model's β was significantly greater than that of the double-update model's (paired $t(10.2) = 4.4$, $p = 0.001$), and again, the targeted-update model's β was significantly greater than that of the double-update model's (paired $t(13.9) = -4.5$, $p < 0.001$). When comparing between model initialization protocols, the single-update model's estimated β was significantly greater under the alternative initialization protocol (paired $t(16.9) = -2.6$, $p = 0.02$), but no differences were observed for the double-update or targeted-update models.

Finally, it was observed that the single-update model's estimated initialization value (under the alternative initialization protocol) was significantly less than that of the double-update model's [paired $t(17.5) = -2.3$, $p = 0.04$], but there was no difference between the initialization values of the double-update and targeted-update models. The fact that the "initVal" parameter was estimated as fairly negative across the three models indicates that our participants were less likely to perform actions that they had been instructed were not advantageous.

The apparent flexibility of the α and β parameters in the presence of additional update mechanisms and an alternate initialization protocol suggests that the core mechanism of these models (described by equations 1, 2, and 3) is capable of fitting participant data in the presence of (or perhaps in spite of) a number of incidental factors. As a consequence, the effect of well-motivated model features have the potential to be obscured by over-flexibility of more "core" model elements. This adaptability poses concern for researchers who seek to explain behavioral, neural, or other forms of data through this approach. In the process of determining the validity of a model, researchers would be well-served by testing multiple model variants under various starting conditions and examining the relationships between model fits, parameter estimation, and differences between model architectures.

References

Cavanagh, J., Frank, M., Klein, T., & Allen, J. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage*, *49*(4), 3198-3209.

Reinhart, R., & Woodman, G. (2014). Causal control of medialfrontal cortex governs electrophysiological and behavioral indices of performance monitoring and learning. *Journal of Neuroscience*, *34*(12), 4214-4227.

Reiter, A., Koch, S., Schrodinger, E., Hinrichs, H., Heinze, H., Deserno, L., et al. (2016). The feedback-related negativity codes components of abstract inference during reward-based decision-making. *Journal of cognitive neuroscience*, *28*(8), 1127-1138.