Modelling Human Information Processing Limitations in Learning Tasks with Reinforcement Learning

Tyler Malloy (mallot@rpi.edu)

Chris R. Sims (simsc3@rpi.edu)
Department of Cognitive Science
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180

Keywords: Reinforcement Learning; Information Theory; Bounded Rationality; Rational Inattention

Introduction

In behavioral economics, 'rational inattention' (C. A. Sims, 2010) has been proposed as a theory of human decision-making subject to information processing limitations. This theory hypothesizes that decision-makers act so as to optimize a trade-off between the utility of their behavior, and the information processing effort required to reach a good decision. Shannon information has been proposed as a means of quantifying this information processing cost. However, existing models in the rational inattention framework do not account for the learning dynamics that underlie human decision-making. In order to incorporate the impact of cognitive limitations on learning, we extend the traditional reinforcement learning objective to incorporate a bound on the Shannon information of the learned policy (see also Lerch & Sims, 2019). Using experimental data from a previously-studied learning paradigm (Niv et al., 2015), we show that our method can be used to represent differences in participants' performance as resulting in part from utilizing different capacities for storing and processing information.

Rational Inattention

According to theories of rational inattention, human decision-makers seek to maximize the following objective (Jung, Kim, Matějka, & Sims, 2019):

$$\max \mathrm{E}[U(X,Y)] - \lambda I(X,Y), \tag{1}$$

where $U(X,Y)$ describes the utility of choice Y in state X, and $I(X,Y)$ represents the mutual information between the state X and the action Y.

The result of altering the traditional expected utility maximization with a regularization term based on mutual information is a constraint on the information-theoretic complexity of the decision-makers' behavior. This limitation is proportional to the scale of the parameter λ; as λ increases, simpler policies will be preferred over increased expected utility. In the extreme, a decision-maker would act randomly or else choose the same action regardless of his or her state (ignoring all information from the environment).

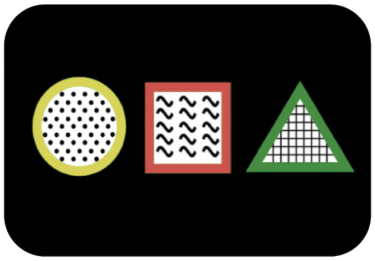

Figure 1: Example of stimuli used in the Niv et al. paradigm. Each of the 9 features was randomly assigned to one of the three possible objects, with no feature present more than once in the same stimulus.

Feature Reinforcement Learning

The specific reinforcement learning algorithm we are interested in extending is a variant of Q-learning defined in (Niv et al., 2015) called Feature Reinforcement Learning (FRL). The algorithm defines the value of an option $V(S)$ in a contextual n-armed bandit learning task to be the sum of the values of the features that make up that option:

$$V(S) = \sum_{f \in S} W(f), \tag{2}$$

where the weights of each feature are updated based on the selection that was made by the participant and the reward that was observed as follows:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \eta[R_t - V(S_{\text{chosen}})] \; \forall f \in S_{\text{chosen}}. \tag{3}$$

FRL was developed to explain human learning performance in domains with high-dimensionality. In their experiment, participants were presented with stimuli varying in color, shape, and texture. Each feature dimension had three possible feature values (for example, stimuli could be red, green, or yellow). The task for participants was to learn which of the nine possible features leads to the highest probability of reward (Figure 1), changing roughly every 20 episodes.

The results shown in (Niv et al., 2015) indicate that it is possible to achieve high predictive accuracy on the selections made by participants using the standard FRL model. In the following section we show that greater predictive accuracy can be achieved by determining the capacity for storing and processing information that is used by each of the participants, and modelling their resulting behaviour with the capacity-limited FRL method.

Capacity-Limited FRL

Applying the learning objective defined in (1) onto the domain of reinforcement learning results in an algorithm that allows us to define a capacity for the amount of information that is used to represent our agent's policy. The two additional hyper-parameters are the capacity-limit C, which is determined for each participant individually using the same method as described in (Niv et al., 2015), as well as the feature weight adjustment learning rate $\alpha = 1e-3$ for all participants.

Algorithm 1: Capacity-Limited FRL

```
Initialize: Feature weights $W(f) = \bar{0}$
Initialize: Hyper-parameters: $\alpha, \beta, \eta, \delta, C$
**for** *each participant selection S* **do**
    Predict choice with probability distribution $\pi(A|S)$
    **for** *each feature f in selection S* **do**
        $W^{\text{new}}(f) = W^{\text{old}}(f) + \eta[R_t - V(S_{\text{chosen}})]$
    **for** *each feature f not in selection S* **do**
        $W^{\text{new}}(f) = (1-\delta)W^{\text{old}}(f)\ \forall f \notin S_{\text{chosen}}$
    **while** $I(\pi(a|s)) > C$ **do**
        **for** *each f in W(f)* **do**
            $f = f - \alpha(f - \sum_{f \in F} W(f)/|W(f)|)$
```

The constraint on the amount of information used to represent performance is determined by the magnitude of the capacity parameter C, which performs the same function as the parameter λ in Eq (1). Decreasing the value of C results in a more and more strict limitation on the amount of information that is used by the model to represent the performance of the participant. The algorithm iteratively updates the RL Q-table to decrease the mutual information until it is below the bound. In the next section, we fit this parameter to each of the individual participants performance in the contextual n-armed bandit learning environment. This algorithm demonstrates that the mutual information regularized expected utility maximization approach that is described in Eq (1) is applicable into the domain of human reinforcement learning.

Results

The original experiment design described in (Niv et al., 2015) includes 2 different speed trials, fast (500ms) and slow (1.5s) response times, with the slow response times used during trials to allow for a fMRI scanner enough time to capture data for a separate analysis that is not discussed further. Hyper-parameters were originally fit by minimizing negative log posterior individually for both the slow and fast trials. However, one potential benefit of the capacity-limited approach is that the information capacity parameter C could be the same across different tasks for the same participant, as long as factors such as motivation and attention remain consistent enough across the different tasks. To support this, we instead fit both models to the entire data set for individual participants using the Python Scipy minimization package, and compare the performance of the FRL and CLRL methods. These results indicate that it is possible to determine the information capacity that is used by a participant in a learning task, even across tasks with slightly different cognitive requirements such as the different time constraints shown here.

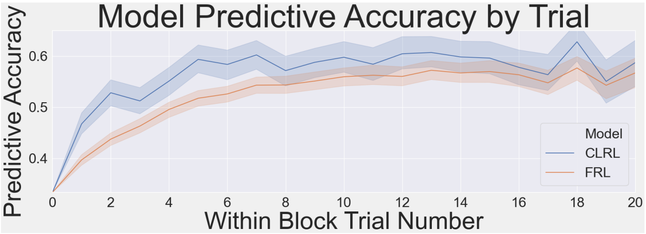

Figure 2: Mean predictive accuracy of CLRL and FRL models based on parameters fit to minimize negative log loss across both fast (500ms) and slow (1.5s) response times. Error bars represent 99% confidence intervals.

The high predictive accuracy of the CLRL model when fit to the entire data set demonstrates a similarity of participant's information processing capacities across different tasks. Although the individual sources of these capacities can be varied, from attention and motivation to differences in cognitive abilities, this model determines the amount of information required to represent participants' learned behaviour. This difference represents one possible explanation for less than optimal performance on learning and decision making tasks that is observed with human participants. By connecting the information-constrained maximum utility with reinforcement learning, this algorithm expands the application into learning tasks. In developing this algorithm, we further support the conceptualization of rational decision makers as Shannon information channels with a limited capacity for storing and processing information that is efficiently allocated to maximize reward when learning and making decisions.

References

Jung, J., Kim, J. H., Matějka, F., & Sims, C. A. (2019). Discrete actions in information-constrained decision problems. *The Review of Economic Studies*, *86*(6), 2643–2667.

Lerch, R. A., & Sims, C. R. (2019). Rate-distortion theory and computationally rational reinforcement learning. *Proceedings of Reinforcement Learning and Decision Making (RLDM) 2019*, 7–10.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neurosci*, *35*(21), 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015

Sims, C. A. (2010). Rational inattention and monetary economics. In *Handbook of monetary economics* (Vol. 3, pp. 155–181). Elsevier.