Timing in Reinforcement Learning Models of Classical Conditioning

Elliot A. Ludvig[1], Richard S. Sutton[1], & E. James Kehoe[2]
[1]University of Alberta and [2]University of New South Wales

The temporal-difference (TD) algorithm from reinforcement learning is a prominent computational model of reward-related learning in the brain and behavior. The equation of the TD (or reward-prediction) error with the phasic responding of midbrain dopamine neurons represents the dominant computational hypothesis for processing by these neurons (Schultz et al., 1997). In this poster, we discuss a three-part computational framework for building and extending current reinforcement-learning models of conditioning. As an illustrative example of this framework, we evaluate how the choice of stimulus representation influences the timing and generalization of reward predictions in these TD models of conditioning.

When animals are exposed to reliable pairings of stimulus and reward, they not only learn the simple contingency between the two stimuli, they also learn the key temporal relationships. For example, when rabbits are trained that a stimulus light is followed 500 ms late by an annoying puff of air to the eye, rabbits learn to blink their eyes in response to the light, with the point of maximal closure occurring around 500 ms after the stimulus (e.g., Smith, 1968; Kehoe et al., 2009). This adaptive timing occurs right from the very first time that animals exhibit a measurable conditioned response (Kehoe et al., 2008). Most computational models of classical conditioning fail to address these real-time features of responding, focusing instead on questions about what the conditions are for learning.

Here, we present a computational framework for extending models of conditioning to include these real-time components of responding. Figure 1 shows how we divide the problem of conditioning into 3 processing stages: stimulus representation, prediction learning, and response generation. This general framework allows for the evaluation of the separable contributions of each of these computational components to the behaviour of the model.

We used this 3-pronged framework to develop a full model of real-time responding in classical conditioning. The model uses the standard TD learning algorithm (Schultz et al., 1997; Sutton, 1988), but

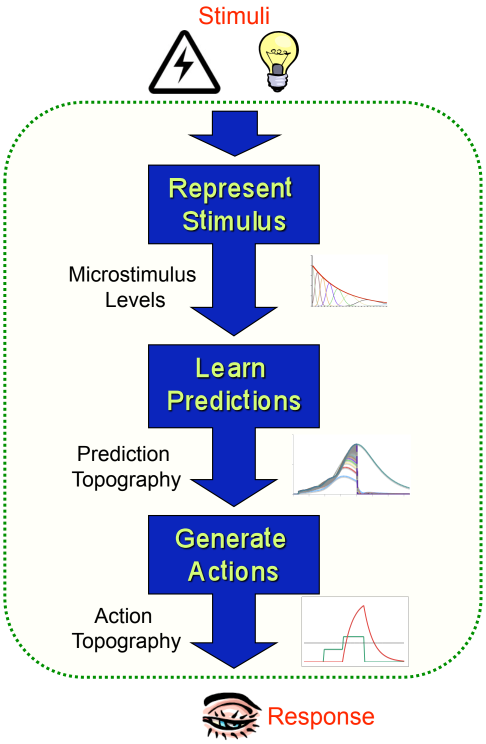

Figure 1. Three stages of information processing in a computational model of classical conditioning.

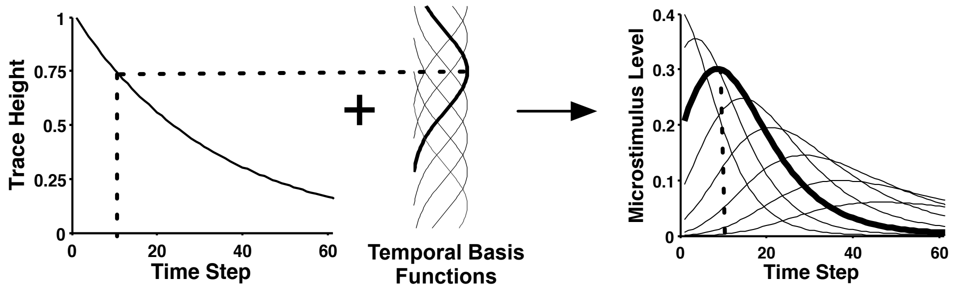

Figure 2. Generation of microstimuli (stimulus features) through the coarse coding of a continuously decaying memory trace. Left panel is the memory trace; middle panel is the coarse coding; right panel is the resultant microstimuli. From Ludvig et al. (2008).

introduces a new temporal stimulus representation. Figure 2 depicts how, in this *microstimulus* representation, all stimuli, including rewards, spawn a series of internal microstimuli that grow weaker and more diffuse over time. These microstimuli are generated by a coarse-coding of a continuously decaying memory trace of that stimulus, through a series of radial basis functions. These microstimuli serve as the features that the TD algorithm uses to predict future reward. This natural stimulus representation produces better correspondence with existent data from the phasic firing of dopamine neurons, especially in experiments where rewards are omitted or mistimed (Ludvig et al., 2008).

In the future, we anticipate following this framework to explore further stimulus representations, new learning algorithms, and alternate response generation mechanism. For example, we are currently examining whether switching to average-reward TD as the learning algorithm allows us to deal with different empirical phenomena, such as the timescale invariance of learning and hyperbolic temporal discounting. Another project evaluates a simple, empirically supported response rule allows for novel predictions about timed responding in rabbit eyeblink conditioning, including the effects of hippocampal lesions on trace conditioning (Ludvig et al., 2009). Our long-term goal is to leverage this three-component framework into developing a suite of reinforcement-learning models of animal decision-making that learn from the real-time flow of experience.

References

Kehoe, E. J., Ludvig, E. A., Dudeney, J. E., Neufeld, J., & Sutton, R. S. (2008). Magnitude and timing of nictitating membrane movements during classical conditioning of the rabbit (*Oryctolagus cuniculus*). *Behavioral Neuroscience, 122*, 471-476.

Kehoe, E. J., Olsen, K. N., Ludvig, E. A., & Sutton, R. S. (2009). Scalar timing varies with response magnitude in classical conditioning of the nictitating membrane response of the rabbit (*Oryctolagus cuniculus*). *Behavioral Neuroscience, 123*, 212-217.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation, 20*, 3034-3054.

Ludvig, E. A., Sutton, R. S., Verbeek, E. L., & Kehoe, E. J. (2009). A computational model of hippocampal function in trace conditioning. *Advances in Neural Information Processing Systems* (*NIPS-08*)*, 21*, 993-1000.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*, 9-44.