

## Lecture 15: Reinforcement learning models

1. We will be brief and selective in reviewing the Niv & Montague chapter (22). This is not because the material isn't important, but because of its nature. It provides an exceptionally well organized historical and conceptual arrangement of the theoretical background to matters we've been discussing over the past few lectures. Read the chapter carefully to help structure your knowledge.
2. We will explicitly highlight a few points we hadn't previously encountered. The first is that TD learning is a variety of *reinforcement learning* (RL), and RL isn't the only kind of learning model. There are also families of models called 'supervised' and 'unsupervised'. The brain may involve some learning processes that would be better described by these types of models.
3. RL is a *normative* framework. That is, it rests on the assumption that the learning system is evolved to optimize some function. We've already encountered some candidate functions that have featured in neuroeconomic models: algorithms for maximizing EV and for minimizing risk. The extent to which real brain systems actually optimize in their performance must be determined by comparison of predictions based on the normatively derived functions with empirical data. This is, of course, standard operating procedure in economics.
4. We've seen how TD learning can allow a system to predict the timing of rewards, and the rate of reward within an interval. But animals must also learn to select behaviors. When they can estimate the comparative EVs of outcome states associated with alternative actions, then TD learning can suffice: the animal should simply estimate the EVs of outcomes and take whatever

action is linked to the outcome in question. But what might a system do when it lacks a model that relates its possible actions to outcomes?

5. A standard response to this problem in Artificial Intelligence (AI) is to build an *actor-critic model*. Such a model comprises two units, where the units can be computers of any level of complexity, but might be as simple as individual neurons. One unit, the adaptive critic element (ACE) implements TD learning to estimate values of states of the environment. These estimates are used to train the second unit, the associative search element (ASE), which learns to optimize actions by trial-and-error. (Note that for a typical animal in a typical environment, this will be very dangerous unless the animal can test the consequences of its actions in *simulations* that its brain can run. This may be the basic source of evolutionary pressure for the development of cognition.) Note that in some environments actor-critic models will fail to converge, so that an animal implementing the model will not be well modeled as an economic agent. In other environments, actor-critic models can get stuck in local maxima (sub-optimal equilibria).
6. Another approach, *Q-learning*, is a modified version of TD learning in which the system learns the values of specific state-action pairs, rather than the values of states only. The learning rule is

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta \delta(t).$$

There are two forms of the TD prediction-error term. The first form

$$\delta(t) = r_t + \max_a \gamma Q(S_{t+1}, a) - Q(S_t, a_t)$$

compares what happens to the best action that was possible. This is 'off policy' learning, and so presumes capacity for simulation. The alternative form of the TD prediction-error term,

$$\delta(t) = r_t + \gamma Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)$$

compares the outcome with the action actually chosen. This is an efficient way to learn in an environment that isn't very dangerous.

7. It has been hypothesized that dopamine signals in the reward circuit implement the critic in an actor-critic model, while a circuit that projects from SNPC to dorsal striatal areas implements the actor. Evidence to specifically support this hypothesis has not yet been forthcoming.
8. RL models have also been applied to animals' choices of rates of behavioral response. The net rate of reward partly determines (along with relative energy costs) the opportunity cost of a given behavioral policy for an animal. Consider two alternative policies A and B, and suppose that A is chosen over B. The higher the reward rate from B, the more vigorously A should be pursued by an economically rational animal.
9. Niv and Montague hypothesize that tonic dopamine levels represent associations between net rates of reward and environments. The evidence to date for this hypothesis comes from biochemical manipulations and studies of damaged and diseased brains.
10. Yu & Dayan have suggested that high acetylcholine levels in striatum signal *expected uncertainty*, known variability in the environment, while high norepinephrine levels signal *unexpected uncertainty*, unknown variability. If this is right,

then prediction errors in phasic dopamine signals will carry different implications depending on the acetylcholine / norepinephrine ratio. Where acetylcholine dominates, prediction errors indicate unexpected changes in the environment. Where norepinephrine dominates, prediction errors indicate higher expected variability.

11. It's been suggested that animals should be encouraged to explore their environments by finding novelty rewarding in itself. This would be implemented by a learning algorithm that optimistically took novelty as predictive of reward, that is, that satisfied  $r_{new}(t) = r_t + novelty(S_t)$ . Ng *et al* show that this does not significantly change the qualitative learning properties of a TD learning system.
12. RL works efficiently when animals parse the world in terms of the RL models. Unsurprisingly, they tend to perform very badly when representational fits between perceptual spaces and learning models are poor. Are natural learning systems' competence zones restricted to environments to which RL model elements have been pre-adapted by natural selection, or are these elements relatively plastic? (Clearly, at least, humans can adjust them using technology.)