

Lecture 17: Implementation of the training of stochastic neural responses

1. Consider the model of learned addiction we considered in the previous lecture. Note that it does not explain addiction by hypothesizing a missing capacity in the learning circuit. If a TD learning implementation is supplemented by a PV mechanism, then the circuit's capacity for optimal learning is limited only by its access to information about the environment. Addiction is explained as a consequence of the system's working *properly* in environments – e.g., casinos – designed to exploit its relative disconnection from wider system goals on longer timescales (e.g., paying off a mortgage). Such environments didn't exist when the brain evolved, so selection didn't favour brains that could cope with them.
2. I stress the point above to emphasize that the models of neural learning we've examined so far, despite their partial basis in neural observation, are still basically derived from the logic of economics: we abstractly characterize a valuation problem, then ask what kind of computer of values could solve it. (In cognitive science, this method is explicitly associated with Chapter 1 of David Marr's posthumous classic *Vision* [1980].) Saying, as we have, that observations are 'consistent with' the hypothesis that ventral striatum 'implements' TD learning and PV is still a very long way from offering a directly testable claim about a neural mechanism. The material reviewed in the chapter by Lee and Wang takes us a *bit* closer to this.
3. We focus on just one, quite general, aspect of the implementation problem. Many learned system responses aren't deterministic. Where they are, systems tend to get 'locked' into types of environments. No doubt primates *are*

environmentally locked to some extent, but their distinctive evolutionary strategies are built around cognitive plasticity. So we face the challenge of explaining how learning affects neural mechanisms in such a way that response *probabilities* vary with changes in objective expected values of environmental contingencies.

4. One obvious kind of test case for this is learning in games where all NE are in mixed strategies. Such learning must obviously consist in some sort of ‘tuning’ of stochastic responses. Matching pennies is a game of this sort. To remind you, here is an instance of Matching Pennies:

	H	T
H	-2, 2	2, -2
T	2, -2	-2, 2

NE = (randomize, randomize)

5. Lee’s group investigated the performance of non-human primates in Matching Pennies by having monkeys play against computers running three different algorithms. Algorithm 0 plays the NE strategy unilaterally, without looking for patterns to be exploited in the monkeys’ responses. In such cases monkeys show strong biases toward one or another of the pure strategies. (The sources of these biases in the experiments is unknown and was not investigated.) Algorithm 1 detects such biases and counteracts them by driving the computer toward the opposite bias. Monkeys respond by dynamically conditioning switches between pure strategies on specific wins and losses. This ‘win-stay, lose-switch’ (WSLS) strategy is an implementation of matching behaviour. It is as good a

- strategy as any other against Algorithm 1. Finally, Algorithm 2 looks for patterns in recent histories of monkey play compared with the monkey's gains and losses and exploits any such patterns. The only NE response to Algorithm 2 is randomization between pure strategies. Monkeys learn this.
6. A similar procedure was followed with Rock-Paper-Scissors as the game. In response to Algorithm 0 monkeys tended to settle on a preferred pure strategy. In response to Algorithm 1 they played the Cournot best response, biasing their choices in each round  $n$  by reference to the pure strategy that would have won or did win in round  $n - 1$ . Monkey response to Algorithm 2 was especially interesting in R-P-S. They *approximated* but didn't quite achieve randomization. In particular, they came as close to randomization as a TD learning rule can get. (Note that possible availability of a PV mechanism is irrelevant here, since as the best response approaches randomization, the value of the predictor approaches 0.)
  7. So now we go looking for a neural mechanism that will adjust stochastic response frequencies toward WSLS when the environment strategically adapts to it but doesn't model it, and implements reinforcement learning when the environment strategically models it.
  8. Individual neurons have been identified in dlPFC that modulate their activity in ways that allow for trial-by-trial comparison of two alternatives, as in Matching Pennies. Crucially, some dlPFC neurons store 'eligibility traces', that is, modulate their response probabilities in light of results of previous responses. This would be necessary in a neural system able to learn best (or near-best) replies to Algorithm 2. (Note that dlPFC receives signals from VS.)

9. Individual neurons in ACC were also studied. Their activity co-varied with calculated reward prediction errors, but not with changes in value functions.
10. Suppose that some individual dlPFC neurons adjust their stochastic response probabilities in such a way as to implement WSLS when randomization doesn't improve on it, and randomize when the environment responds as if strategically modeling the organism's behaviour. How might they do this? Lee and Wang review several models.
11. The first model is a specific type of 'ramping-to-threshold' model, called a 'drift-diffusion' model. Suppose we have two alternatives  $X_1$  and  $X_2$  for comparison, where  $X = X_1 - X_2$ . Then the dynamics of  $X$  are modeled by drift diffusion if

$$dX/dt = \mu + \omega(t)$$

where  $\mu$  is the drift rate and  $\omega(t)$  is a white noise of zero mean and standard deviation  $\sigma$ .  $\mu$  represents a bias in favour of one alternative  $X_1$  or  $X_2$ . The system is a perfect integrator of the input

$$X(t) = \mu t + \int \omega(t') dt'$$

and terminates whenever  $X(t)$  reaches a positive threshold  $\theta$  (choice 1) or  $-\theta$  (choice 2). If  $\mu$  is positive then choice 1 is correct, while choice 2 is an error; otherwise the opposite. If  $\mu$  is 0 the system is 'set' to randomize.

12. Can neural circuits implement this model? They can't perfectly integrate inputs, because they 'leak' – that is, drift with time independently of  $\mu$ . (Put in plain terms: they forget.) However, it has been demonstrated that this can be corrected by recurrent activation. So as long as input meets some persistence threshold, and the neuron is embedded in a network that receives stabilizing feedback from elsewhere in

the brain (i.e., there's recurrence), then a neuron's stochastic response rate can systematically adjust in the way approximately described by the drift diffusion model.

13. We can get less abstract in search of greater biophysical accuracy. A *spiking network* model takes account of the specific rise-times and decay-times of synapses. Lee and Wang point out that “synaptic dynamics turn out to be a crucial factor in determining the integration time of a neural circuit dedicated to decision making, as well as controlling the stability of a strongly recurrent network.” Lee & Wang don't provide a formal example of a spiking network model, presumably because there are too many possible variations. For a tour of these, see Daniel Amit, *Modeling Brain Function*, Cambridge University Press 1989.
14. In general, what hold stochastic neural firing frequencies within stable but adjustable bands are network properties that create complex dynamics with multiple attractors. For an introduction to such systems, see Mitchel Resnick, *Turtles, Termites and Traffic Jams* (MIT Press, 1997). Pairs of neural networks linked to one another by both excitatory and inhibitory connections decide between two alternatives A and B, where  $C_A$  and  $C_B$  denote recurrence of input favouring A and B respectively, according to the softmax function

$$P_A(C_A - C_B) = 1 / (1 + \exp(-(C_A - C_B)) / \sigma)$$

where  $\sigma$  denotes the ‘extent of’ stochasticity’ in the network. Such a system will decide between A and B even when absolute magnitudes of  $C_A$  and  $C_B$  are small, and when  $C_A = C_B$ . The function performs well in describing monkey behaviour in the Matching Pennies game. The model hasn't been tested on single neuron responses in dlPFC (or elsewhere) during game play. However, it has produced good fits in estimating firing rates in area LIP when monkeys

learned to anticipate stochastic changes in visual displays. (You should recall area LIP from our discussion of the early Glimcher inspection game experiments.)

15. If neurons ramp to threshold, then learning must involve adjustments to the thresholds (e.g., to  $\theta$  in the drift diffusion model). Lee & Wang suggest that this is what the striatal dopamine signal does. Wouldn't that tie everything up neatly? It certainly is a natural hypothesis, given the model. Lee & Wang don't indicate their specific evidence for it, however.
16. Lee and Wang used their model to generate a simulation of neural learning of the monkeys' task in Matching Pennies. It captures the broad characteristics of the observed behaviour, though it under-predicts the monkeys' frequency of use of WSLS against Algorithm 1. To reproduce this, they adjust the model by using "a different learning rule, according to which synapses onto both neural populations (selective for the chosen and unchosen targets) are modified in each trial. This is akin to a 'belief-dependent learning rule'." I don't find this entirely satisfying. It doesn't say enough to allow us to judge how substantial is the change in the learning rule.
17. Note that the model used for the simulation required different parameters for play against each of the three algorithms. The simulation can't be regarded as a full test of the hypothesis if these must be adjusted by hand. Thus Lee & Wang "incorporated a meta-learning rule proposed by Schweighofer and Doya (2003) that maximizes long-term rewards." Thus the simulation really tests a *joint* hypothesis: that some neurons (somewhere in the brain) tune their response probabilities by ramping to thresholds via strongly recurrent dynamics, thereby learning strategically superior stochastic behaviours, *and* parameters that govern applications of this learning model are themselves learned by

some so far unspecified neural process somewhere else in the brain that implements the Schweighofer-Doya meta-learning rule.

18. Lee & Wang never say how to generalize the softmax function that describes the dynamic attraction so as to incorporate choices among more than two alternatives. Nor do they mention any test of a relationship between their model and the reported pattern of monkey play in R-P-S. I don't know why they open their chapter by describing the R-P-S experiment and observations, and then never again bring it up, or mention the evident requirement it raises for a model of neuronal response that can explain learning to mix three pure strategies.