Introduction

These notes are about reinforcement learning. I offer an error bounty of between 20p and 2 pounds for mistakes. Contact me at `conor.houghton@bristol.ac.uk` or come up after a lecture.

Action selection

Classical conditioning tests responses using an experimental paradigm which in its reliability differs from the real world. The bell rings and the dog gets fed, little in our lives offers this sort of reliability, we hear a bell and we know we need to leave the building and can guess the probability that there is a fire or that we'll get into trouble for not taking part in a drill; we evaluate based on probabilities and on estimated likelihoods of various future harms and rewards. Action selection relates to the problem of choosing actions in this uncertain and probabilistic world. The basal ganglia is considered the seat of action selection.

The classic experiment in action selection is described in [1]; it involves one-arm-bandits, modelled on the gambling machines you find in amusement arcades and casinos. The player is presented with a choice of four one-arm-bandits and chooses one, they play and receive a reward; they then choose again. The mean rewards are adjusted over time. The question is, how does the player choose which bandit to play and how to the make sure to continue exploring the other bandits in case they become better as the means are changed.

One strategy is to estimate rewards and match choice probability to these estimates. Hence, say m_i is the estimated reward for action i; then the probability of choosing i is

$$p_i = \frac{\exp \beta m_i}{\sum_j \exp \beta m_j} \tag{1}$$

where β is a 'choosiness' or exploration parameter. If $\beta = 0$ then all actions are chosen with equal probability and there is little discrimination; if β is very large, the action with the highest estimated reward is chosen with probability near one, and there is little exploration. Of course m_i is updated based on outcome,

$$m_i \to m_i + \eta\delta \tag{2}$$

where η is a learning rate and

$$\delta = r - m_i \tag{3}$$

is the error in the expected reward.

Basal ganglia

The basal ganglia, Fig. 1, is a collection of sub-cortical brain areas, or nuclei, found near the center of the brain and connected to the cortex, thalamus and brain stem. It is thought to be important in decision making, action selection and in the regulation of some routine behaviors, like eye movements. Its constituent nuclei include the striatum, the globus pallidus, the substantia nigra, the nucleus accumbens, and the subthalamic nucleus. These are all quiet complex with different, sometimes very distinctive, neurons; they are frequently divided still further, so, for example, in models of basal ganglia the globus pallidus is divided to the excitatory part of globus pallidus and the inhibitory. There are also channels in the basal ganglia, there are parts of each nucleus that correspond to different muscles and different parts of the body.

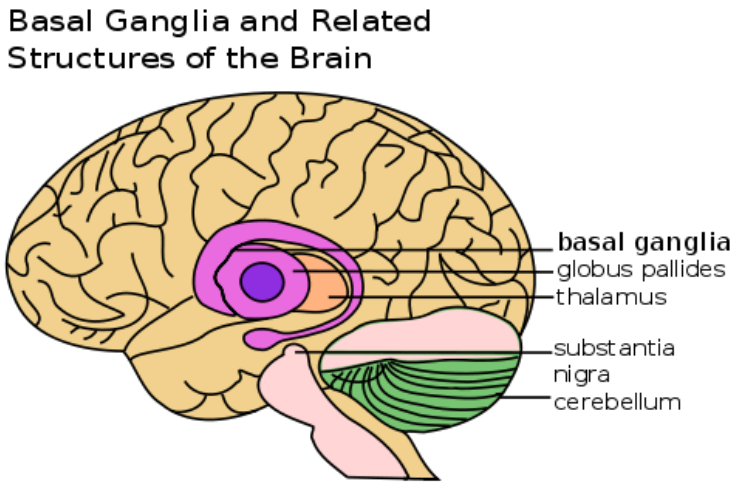

Figure 1: The basal ganglia. This picture is slightly misleading since it looks like the basal ganglia is on the surface of the brain when it is on the inside, it is as if the beige parts of the picture, showing the cortex, are transparent. It also labels the globus pallides and substantia nigra separately as if they were distinct from the basal ganglia, when they are part of it. [Picture from wikipedia]

The basal ganglia has lots of inhibitory projections, so it is thought it acts mostly switching off inhibition. In other words, there are cells which inhibit muscles and the basal ganglia is thought to work by, in turn, inhibiting those cells, see Fig. 2 for a sketch of this mechanism.

Model of action selection in the basal ganglia

This circuit is very similar to the one introduced in classical conditioning, however, although dopamine plays a similar role, describing the error, a different set of dopaminergic neurons may be involved, the dopaminergic neurons of the substantia nigra (SN). The model is represented in Fig. 3.

Evidence for this model is provided by recordings from human patients in [2]; recordings where taken from Parkinson's patients during surgery for deep brain stimulation. The subjects played a one-armed-bandit type game, this version involved drawing cards from two decks, one had a 0.35 chance of reward, the other 0.65. As shown in Fig. 4 there is increased activity in SN when there is a reward from the 0.35 card and a decrease in activity when there is no reward from the 0.65 card.

Very direct evidence is provided in [3]. A clever two phase experiment is used to measure changes in synaptic weight directly. In the first phase SN neurons are stimulated electrically in rat during a lever press task; the optimal stimulation for learning is found; during the second phase the rat is anaesthetised and the cortex, striatum and SN are stimulated directly and the synaptic weights are measured. It is found that the cortico-striatal weights increase during simulated learning.

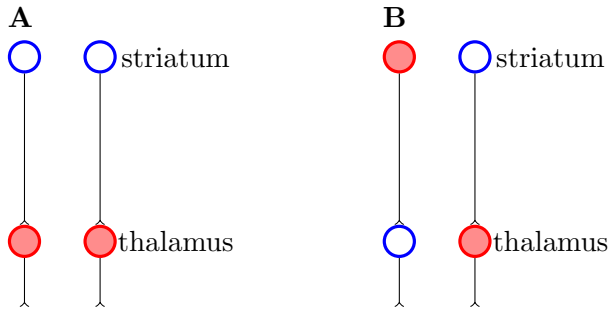

Figure 2: The basal ganglia controls by disinhibition. In **A** the thalamic neurons inhibit the muscles; however in **B** one of the basal ganglia neurons has become active, it inhibits a thalamic neuron, disinhibiting the left muscles.

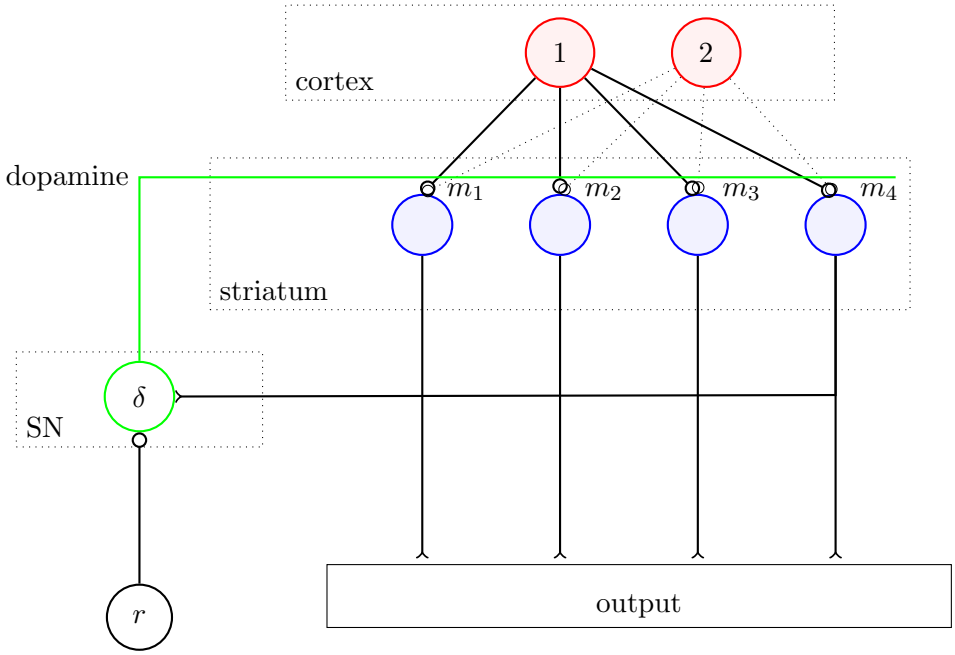

Figure 3: A diagram of the action selection circuit. The cortical input, labeled '1' excites some neurons in the striatum; the synapses have strengths corresponding to the estimated rewards and, in this simple model, this determines the amount of activity in the post-synaptic neuron, these neurons in turn disinhibit neurons in the output, allowing the corresponding muscles to move, the difference between this activity and the reward is represented by dopamine and causes synaptic changes at the cortico-striatal synapses.

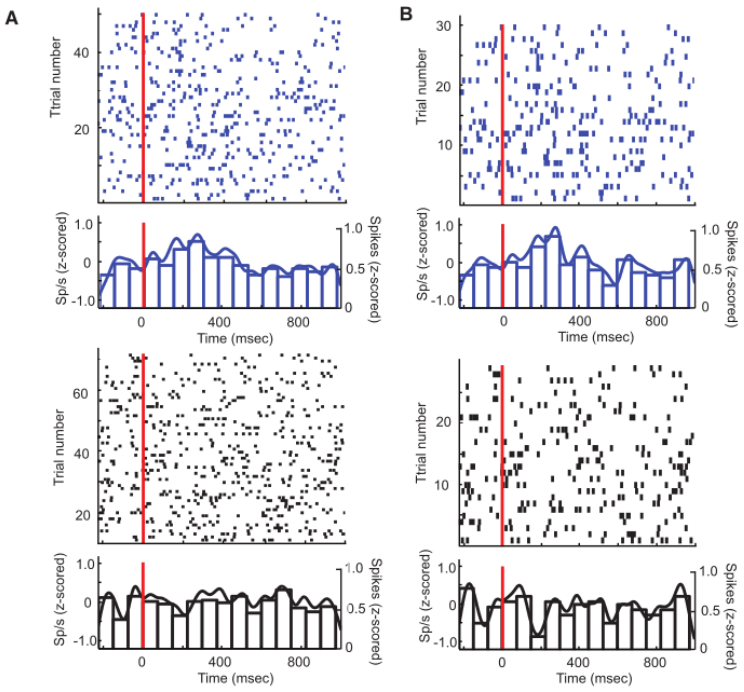

Figure 4: Response from SN in one patient during a game of chance from [2], The top, blue, graphs shows increased activity during an expected reward, the bottom, black, shows decreased when a reward is expected, but doesn't appear.

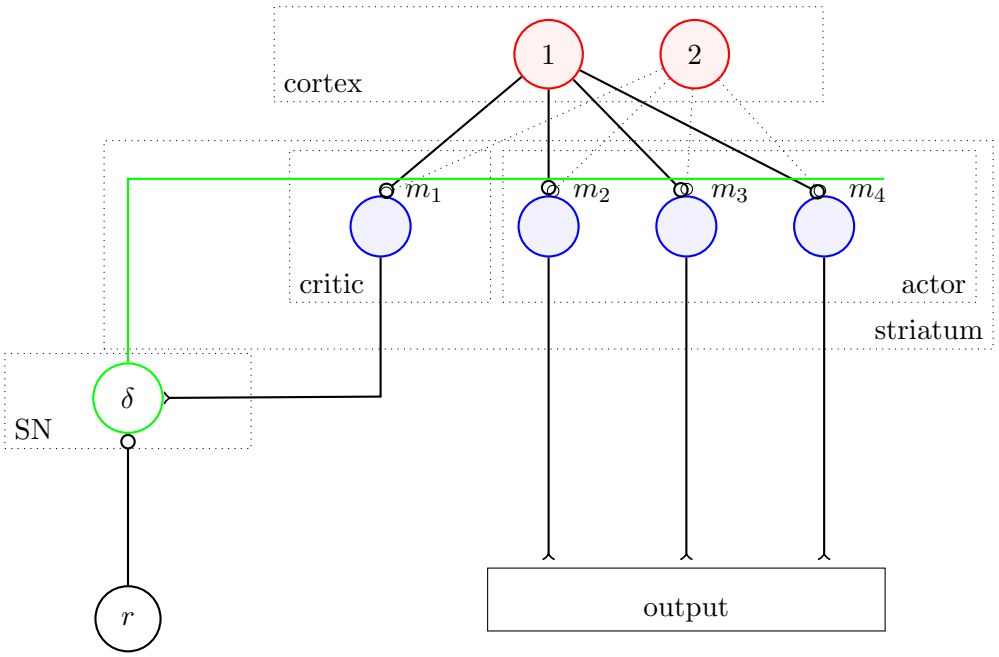

Figure 5: A diagram of the actor-critic circuit: there is a separate pathway for directing, that is disinhibiting, action and for calculating error.

Actor-critic model

The is plenty of evidence for a SN dopaminergic error signal. However, the connectivity the model involves is not consistent with the connectivity actually observed. The actor-critic fixes this; it divides the striatum into two parts, the large 'actor' part that guides action and is associated with the dorsal, that is front, part of the striatum and a smaller 'critic' part used to calculate reward; this is associated with the ventral, that is lower, part. This circuit is sketched in Fig. 5. In the model actions are selected according to the expected reward m_i, an error is calculated using $\delta = r - w$ and both w and m_i are updated

$$m_i \to m_i + \eta\delta \tag{4}$$

and

$$w \to w + \eta\delta. \tag{5}$$

There is some proof for this model provided in [4]. In this paper subjects play a one-armed-bandit-style game under two conditions, one where the computer chooses, one where the subject chooses. In both the ventral striatum is engaged, but the dorsal striatum is only engaged when the subject chooses.

Rewards and harms

One problem with these models is that all the projections from the cortex to the striatum are excitatory, so all the m_i must be positive; this doesn't reflect the state of the world, there are negative as well as positive rewards and it would seem likely the model should be able to deal with this.

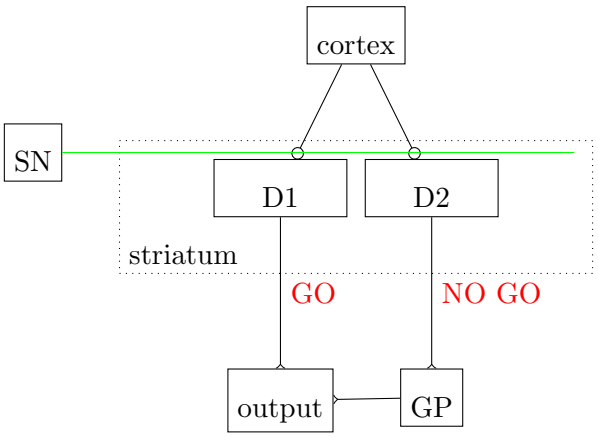

Figure 6: Schematic of the go / no-go circuit; the connect from SN through the synapses is a dopamine error signal.

It is suggested in [5] that the globus pallidus (GP) creates a 'no go' pathway which complements the normal 'go' pathway. In their model there are two different neuron populations in striatum, with different types of dopamine receptors. One population, D1, corresponds to the reward learning we have been looking at so far, the other population, the D2 cells, are involved in learning from harms. These neurons don't project to the output directly, instead the project to GP which, in turn, has inhibitory projectoins to the output; because of the extra layer there is an extra minus sign along this pathway, so it prevents action instead of allowing it.

In this model the net estimated reward is encoded in the different between the D1 and D2 activity, so

$$m_i = w_i^1 - w_i^2 \tag{6}$$

where w_i^1 and w_i^2 are the activity levels, or in this simple model, equivalently, the synapse strengths, for equivalent D1 and D2 neurons. Although both w_i^1 and w_i^2 are positive the extra minus, roughly the extra synapse at GP, allows m_i to be positive or negative. To learn this, the model suggests the D1 and D2 cells react differently to dopamine so that

$$w_i^1 \to w_i^1 + \eta\delta \tag{7}$$

as before, but

$$w_i^2 \to w_i^2 - \eta\delta \tag{8}$$

This is not completely outrageous, it is known that dopamine can have opposite effects on different synapses depending on their receptors, this is the sort of complicated dynamics that neuromodulators support.

This model predicts that a change in dopamine level will have different effects on different circuits, it will increase the strength of the go circuit but decrease the strength of the no-go circuit and the other way around if the dopamine level is decreased. Hence, learning from rewards should be facilitated in people with increased dopamine and learning from punishment from people with decreased dopamine.

This was tested in [5] using subjects with Parkinson's disease. People suffering from Parkinson's disease have a decreased level of dopamine because dopaminergic neurons in SN have

died; one of the main therapies is to give medication which increases the dopamine level. This makes it possible to manipulate the dopamine levels of Parkinson's disease by adjust the level of medication; obviously this relies on the generosity of Parkinson's suffers in participating in experiments which will involve a temporary increase in the severity of their symptoms as their medication is withdrawn. This has been tested using a one-armed-bandit style game, with happy-face pictures acting as rewards and frowny-face pictures as punishments; the predicted effect is observed.

References

[1] Daw ND, O'Doherty JP, Dayan P, Seymour B and Dolan RJ. (2006) Cortical substrates for exploratory decisions in humans. Nature, 441: 876–879.

[2] Zaghloul KA, Blanco JA, Weidemann CT, McGill K, Jaggi JL, Baltuch GH and Kahana MJ. (2009) Human substantia nigra neurons encode unexpected financial rewards. Science, 323: 1496–1499.

[3] Reynolds JN, Hyland BI and Wickens JR. (2001) A cellular mechanism of reward-related learning. Nature, 413: 67–70.

[4] O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K and Dolan RJ. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science, 304: 452–454.

[5] Frank MJ, Seeberger LC and O'Reilly RC. (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science, 306: 1940–1943.