RL 19: Biological and neural RL
(possibly interesting, but not examinable)

Michael Herrmann

University of Edinburgh, School of Informatics

(bonus lecture)
Different types of learning in neural systems
Neural correlates of reward, value and action selection
Computational model provide interpretation of neural processes as reinforcement learning by computation
Inspired by behaviorist psychology, reinforcement learning is an area of machine learning in computer science, concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward.

- Choose actions to move to states which are as good as possible
- Quality of states is measured by the expected future discounted reward
- Expectation is taken w.r.t. to a fixed policy
RL shares with behaviourism

- its origins in animal learning theory
- its focus on the interface with the environment
- states and actions (or: stimuli and responses)
- the idea that *representations* aren’t needed to define optimality

In the end it all comes down to the actions taken and the states perceived.

RL *of course* is all about the algorithms and processes going on inside the agent.

For example, RL (in ML) often considers the construction of internal models of the environment within the agent, which is *far* outside the scope of behaviourism

adapted from http://webdocs.cs.ualberta.ca/~sutton/RL-FAQ.html#behaviorism, emphasis changed
Historical roots: The law of effect

“Connectionism” (E. Thorndike, 1911):

- “satisfying state of affairs” leads to reinforcement of the association between action and this state
- “annoying state of affairs“ leads to weakening of the association between action and this state

Remarks:

- Consequences of behaviour determine what is learnt and what is not
- Thorndike introduced animal studies for verifying predictions made from his theory. He also was among the first to apply psychological principles in the area of teaching (active learning)
- Connectionism implies modelling of higher brain functions as the emergent processes of interconnected networks of simple units. Thorndike provided the first working model.

http://www.lifecircles-inc.com/Learningtheories/behaviorism/Thorndike.html
Learning

Psychology

- **Non-associative learning**: single stimulus *(habituation or sensitisation)*

- **Associative learning**
  - two stimuli *(classical conditioning)*: A neutral stimulus causes a response. After learning, a conditioned stimulus causes a similar response *(unsupervised or Pavlovian learning)*
  - stimulus-response *(operant conditioning, reinforcement learning)*

Machine learning

- Unsupervised learning
- Supervised learning
- Reinforcement learning
Classical condition: Rescorla & Wagner (1972)

Two assumptions:

- learning is driven by error (formalise notion of surprise)
- summations of predictors is linear

Change in value is proportional to the difference between actual and predicted outcome

$$\Delta V_{X}^{n+1} = \alpha_X \beta (\lambda - V_{tot})$$

$\Delta V_X$ change in strength of association of (CS) $X$

$$V_{X}^{n+1} = V_{X}^{n} + \Delta V_{X}^{n+1}$$

US: unconditioned stimulus, CS: conditioned stimulus

$\alpha_X \in [0, 1]$ salience of CS, $\beta \in [0, 1]$ rate parameter (given the US)

$\lambda$ is the maximum conditioning possible (given the US)

$V_{tot}$ is the total associative strength of all CS ($= V_{X}^{n}$ for one CS)


bonus lecture in 2015 Michael Herrmann RL 19
Instrumental conditioning

Thorndike “Animal intelligence: an experimental study of the associative processes in animals” (PhD thesis)
Tested hungry cats in “puzzle boxes”
Definition for learning: Time to escape
Gradual learning curves, did not look like ‘insight’ but rather trial and error

www.princeton.edu/~yael/
Learning paradigms and the brain

- **Cerebellum**: Supervised learning
- **Basal ganglia**: Reinforcement learning
- **Cerebral cortex**: Unsupervised learning

Doya, Kenji. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?. Neural networks 12.7 (1999): 961-974.
Supervised and unsupervised learning

Cerebellar circuit. GC, granule cells; PC, Purkinje cells; CN, deep cerebellar nuclei; IO, inferior olive;
- ○ excitatory connection; • inhibitory connection

Cortical circuits. P, pyramidal neurons; S, spiny stellate neurons; I, inh. interneurons;

In contrast to cerebral cortex and cerebellum, the basal ganglia are structurally and functionally heterogeneous, complex and only partially understood.

Functions:
- Selection and processing of behavioural patterns
- Suppression and inhibition of undesired activation patterns (Gating theory)

Responsible for diseases such as Parkinson’s disease or Tourette syndrome

GPe: globus pallidus external
GPi: globus pallidus internal
STN: subthalamic nucleus
SNc: substantia nigra compacta
SNr: substantia nigra reticulata
Pathways: excitatory (glutamatergic, red), inhibitory (GABAergic, blue), modulatory (dopaminergic, magenta).
Unsupervised, supervised and reinforcement learning

Diagram of neural circuit of the basal ganglia

SNC: substantia nigra, pars compacta; SNr: substantia nigra, pars reticulata
GPI: globus pallidus, internal segment; GPE: globus pallidus, external segment
STN, subthalamic nucleus; ○ excitatory connection; ● inhibitory connection
Unsupervised, supervised and reinforcement learning

Model-free, stochastic action selection (Cortex and basal ganglia)
Unsupervised, supervised and reinforcement learning

Action selection with a forward model.
Differential model-based action selection.

Compensation of sensory feedback delay with a forward model. The thick arrow represents either of the state-to-action mapping by the architectures shown above.
Function of Doya’s brain model

Simulation of behaviour using a forward model instead of the real environment.
Encapsulation of complex decision process in a simple reactive mapping.

Learning (according to F. Wörgötter)

Machine Learning
Anticipatory Control of Actions and Prediction of Values

Reinforcement Learning
Example based

Q-Learning
Monte Carlo Control
Dynamic Prog. (Bellman Eq.)

Classical Conditioning

Q-learning

TD(\lambda)
often \lambda=0

\delta\text{-Rule}
supervised L.

Eligibility Traces

Rescorla/Wagner

Neur.TD-formalism

Neur.TD-Models
("Critic")

ISO-Learning

ISO-Model of STDP

Biophys. of Syn. Plasticity
Glutamate

Differential Hebb-Rule ("slow")

Differential Hebb-Rule ("fast")

STDP

LTP (LTD=anti)

Synaptic Plasticity
Correlation of Signals

Un-Supervised Learning
Correlation based

Hebb-Rule

STDP-Models
Biophysical & network

NON-EVALUATIVE FEEDBACK (Correlations)

EVALUATIVE FEEDBACK (Rewards)
RL: The Computational Neuroscience Approach

- The problem: optimal prediction of future reward
- The algorithm: temporal difference learning
- Neural implementation: does the brain use TD learning?

David Marr (according to P. Dayan)
Reinforcement learning has revolutionised our understanding of learning in the brain in the last 20 years (Y. Niv)

Certain cortical neurons exhibit an elevated (anticipatory) activity during periods before significant behavioural events

Dopamine neurons in the midbrain, however show activities that can be interpreted as an error signal in the theory of reinforcement-learning.


*Neural Computation* 13(4), 841-862.
Reward system in the brain (beyond algorithms)

Overview of reward structures in the human brain. Dopaminergic neurons are located in the midbrain structures substantia nigra (SNC) and the ventral tegmental area (VTA).

Their axons project to the striatum (caudate nucleus, putamen and ventral striatum including nucleus accumbens), the dorsal and ventral prefrontal cortex. Additional brain structures influenced by reward include the supplementary motor area in the frontal lobe, the rhinal cortex in the temporal lobe, the pallidum and subthalamic nucleus in the basal ganglia, and a few others.

Dopamine

Dopamine is commonly associated with the reward system of the brain, providing feelings of enjoyment and reinforcement to motivate a person to perform certain activities.

It is released (particularly in areas such as the nucleus accumbens and prefrontal cortex) by rewarding experiences such as food, sex, drugs, and neutral stimuli that become associated with them.

Dopamine is closely associated with reward-seeking behaviours, such as approach, consumption, and addiction. Recent research suggests that the firing of dopaminergic neurons is motivational as a consequence of reward-anticipation.

Dopamine’s role in motivation, desire, and pleasure

Rats were depleted of dopamine by up to 99 percent in the nucleus accumbens and neostriatum using 6-hydroxydopamine. With this large reduction in dopamine, the rats would no longer eat of their own volition. The researchers of this study concluded that the reduction in dopamine did not reduce the rat’s consummatory pleasure, only the desire to eat.


Mutant hyperdopaminergic (increased dopamine) mice show higher "wanting" but not "liking" of sweet rewards. Mice who cannot synthesise dopamine are unable to feed sufficiently to survive more than a few weeks after birth, but will feed normally and survive if administered L-DOPA.


When monkeys are given a highly palatable food, dopamine levels rise, but levels then decline when the palatable food is available for prolonged periods of time and is no longer novel.

en.wikipedia.org/wiki/Dopamine
Prediction error hypothesis
Dopamine neurons in ventral tegmental area (VTA)


No prediction
Reward occurs

Reward predicted
Reward occurs

Reward predicted
No reward occurs

Striatal neurons show ramping activity that precedes a reward (and changes with learning!)

Daw: Ramping from start to food reward

Schultz
Evolution of reinforcement learning in a model

(a) The bee’s neural network controller. (b) The bee’s action function. Probability of reorienting direction of flight as a function of $P(t)$ for different values of parameters $m; b$. (c) The genome sequence of the simulated bee.

How to implement TD in a neuronal way (F. Wörgotter)

Serial-Compound representations $X_1, \ldots, X_n$ for defining an eligibility trace.

Note: $v(t+1) - v(t)$ is acausal (future!). Make it "causal" by using delays.
How to implement TD in a neuronal way (F. Wörgötter)

#1

Predictive Signals

\[
\begin{align*}
\text{Start: } w_0 &= 0 \\
\text{End: } w_0 &= 1
\end{align*}
\]

\[
\delta = v' + r
\]

#2

\[
\begin{align*}
\text{Start: } w_0 &= 1 \\
\text{End: } w_0 &= 1
\end{align*}
\]

Forward shift because of acausal derivative

#3

\[
\begin{align*}
\text{End: } w_0 &= 1 \\
\text{End: } w_0 &= 1
\end{align*}
\]
Biological evidence

- Evidence Accumulation (Gold & Shadlen, 2007)
- Variants: SARSA (Morris et al, 2006)
- Q learning (Roesch et al, 2007)
- Neuromodulation
  - dopamine phasic: prediction error for reward
  - tonic: average reward (vigour)
  - serotonin phasic: prediction error for punishment?
  - acetylcholine: expected uncertainty?
  - norepinephrine unexpected uncertainty; neural interrupt?
Average firing rate of 19 dopaminergic neurons, recorded in rats performing an odor-discrimination task in which one of the odors predicted that a reward would be delivered in a food-well, with some delay.

Color indicates the length of the delay preceding reward delivery from 0.5 to 7 seconds. Activity is aligned on odor onset (left) and food-well entry (right). Note that the response to the (not fully predicted) reward is similar in all trial types (with the earliest rewards perhaps better predicted, and thus accompanied by smaller prediction errors), but the response at the time of the predictive cue depends on the predicted delay of the reward, with longer predicted delays eliciting a smaller dopaminergic response. Adapted from Roesch et al. (2007) by Y. Niv (2009)
(Action -) Value Function Approximation

In order to reduce the temporal credit assignment problem methods have been devised to approximate the value function using so-called features to define an augmented state-action space.

Most commonly one can use large, overlapping feature (like “receptive fields”) and thereby coarse-grain the state space.

![Image of a grid with circles indicating state space and value function approximation]

Black: Regular non-overlapping state space (here 100 states).

Red: Value function approximation using here 17 features, only.

Note: Rigorous convergence proof do in general not anymore exist for Function Approximation systems.

F. Wörgotter
Example: Path-finding in simulated rats

**Goal:** A simulated rat should find a reward in an arena.

This is a **non-regular RL-system**, because

1) Rats prefer **straight runs** (hence states are often “jumped-over” by the simulated rat). Actions do not cover the state space fully.

2) Rats (probably) use their hippocampal **Place-Fields** to learn such task. These place fields have **different sizes** and cover the space in an **overlapping** way. Furthermore, they fire to some degree **stochastically**.

Hence they represent an **Action Value Function Approximation system**.

F. Wörgötter

---

*Bonus lecture in 2015 Michael Herrmann RL 19*
Place field system

Path generation and Learning

Real (left) and generated (right) path examples.

F. Wörgötter
Randomness of synaptic transmission is harnessed by the brain for learning.
Possible if synapses are “hedonistic”, responding to a global reward signal by increasing their probabilities of vesicle release or failure, depending on which action immediately preceded reward.
Hedonistic synapses learn by computing a stochastic approximation to the gradient of the average reward.

A spiking neural network model of actor-critic learning

- Spiking neural network model implements actor-critic temporal-difference learning
- Combines local plasticity rules with a global reward signal
- Network solves a gridworld task with sparse rewards
- Similar equilibrium performance as look-up table variants

Learning curves for gridworld task for fast (black) and slow (grey) policy learning. Average over 15 trials.

Neuronal implementation of the actor-critic architecture. Each state is represented by a pool of 40 neurons, the critic by a group of 20 neurons, and the actor by one neuron for each possible action (4).

The state signal ($s$) consists of a positive DC stimulus from the environment ($E$) to the appropriate group of state neurons. The action signal ($a$) is defined as the first spike emitted by one of the actor neurons after entering a new state. The reward signal ($R$) has a modulatory effect on the state-critic synapses. The action suppression signal consists of a negative DC stimulus to the actor neurons.
Policy learnt by the neuronal network for the gridworld task (see Figure 5, inset). (A–C) Preferred movement direction for each state for different runs of the neuronal implementation. (D) Preferred movement direction for each state averaged over 10 runs.

Possible due to specific learning scheme for state-critic synapses:

- negligibly plastic except for a short time period when the agent has just left the corresponding state
- sensitive to a characteristic dynamic response of the critic neurons, which encoding change in stimulus
- sensitive to a global signal representing the reward.
Challenges (Y. Niv)

- How does learning from one task affect subsequent learning? Responses of dopamine neurons to stimuli not clearly related to reward prediction
  - Generalisation?
  - Novelty bonuses?

- Hierarchical RL: How does an agent learn useful modules?

- Temporal effects:
  - Is the high degree of noise in behavioural timing consistent with the temporal sensitivity displayed by neural prediction error signals?
  - Nature of (subjective) time and temporal order?

- How is unlearning represented?
  - Extinction learning: Pavlov’s dog eventually stopped to drool


ND Daw & K Doya (2006) - The computational neurobiology of learning and reward. Current Opinion in Neurobiology, 6, 199-204


