RL 19: Biological and neural RL

(possibly interesting, but not examinable)

Michael Herrmann

University of Edinburgh, School of Informatics

(bonus lecture)

Overview

- 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

Reinforcement learning

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

en.wikipedia.org/wiki/Reinforcement_learning

- 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

Behaviourism

has been disparaged for focusing exclusively on behaviour, refusing to consider what was going on inside the head of the subject.

- 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

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

ΔV_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)

λ is the maximum conditioning possible (given the US)

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

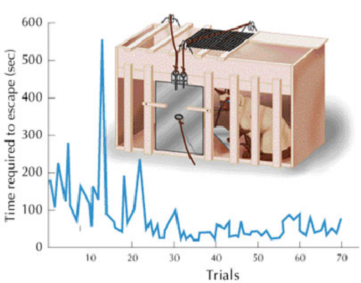
en.wikipedia.org/wiki/Rescorla%E2%80%93Wagner_model, www.princeton.edu/~yael/

Instrumental conditioning

Thorndike "Animal intelligence: an experimental study of the associative processes in animals" (PhD thesis)

Tested hungry cats in "puzzle boxes"

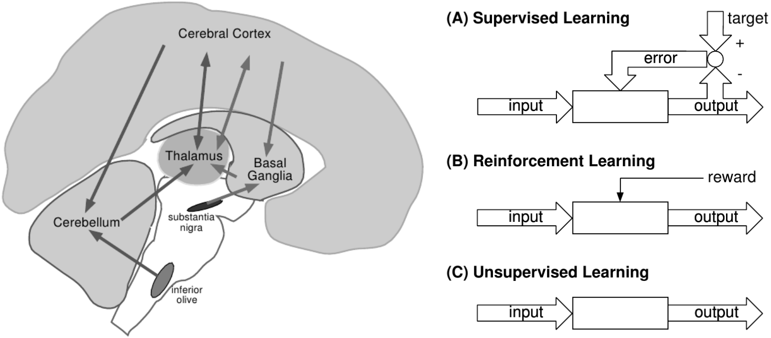
Definition for learning: Time to escape



Gradual learning curves, did not look like 'insight' but rather trial and error

www.princeton.edu/~yael/

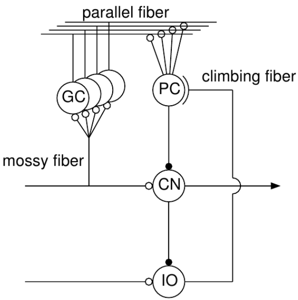
Learning paradigms and the brain

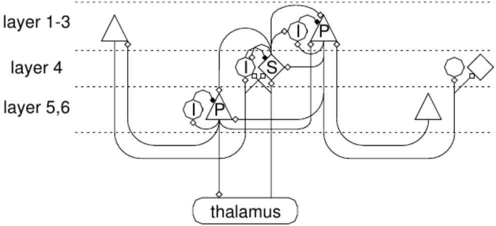


- Cerebellum: Supervised learning
- Basal ganglia: Reinforcement learning
- Cerebral cortex: Unsupervised learning

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;

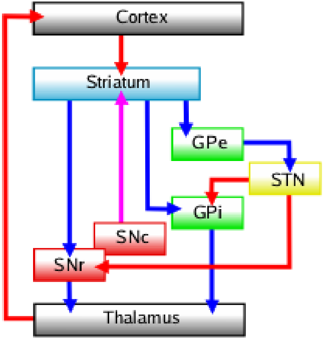
Cortical circuits. P, pyramidal neurons; S, spiny stellate neurons; I, inh. interneurons;

○ excitatory connection; • inhibitory connection

Doya, Kenji. Neural networks 12.7 (1999): 961-974.

Basal ganglia: Reinforcement Learning

- 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

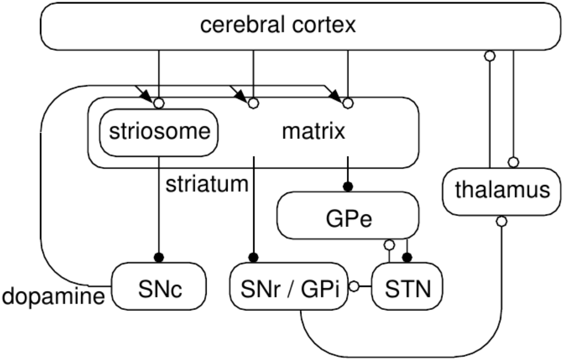
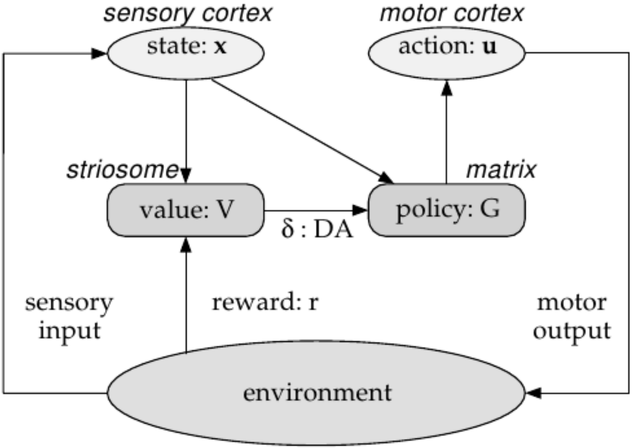


Diagram of neural circuit of the basal ganglia

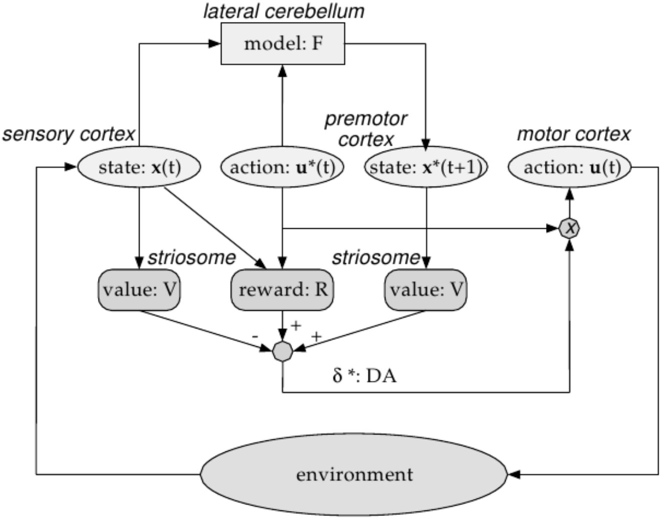
SNc: substantia nigra, pars compacta; SNr: substantia nigra, pars reticulata
GPi: globus pallidus, internal segment; GPe: globus pallidus, external segment
STN, subthalamic nucleus; ○ excitatory connection; • inhibitory connection

Unsupervised, supervised and reinforcement learning



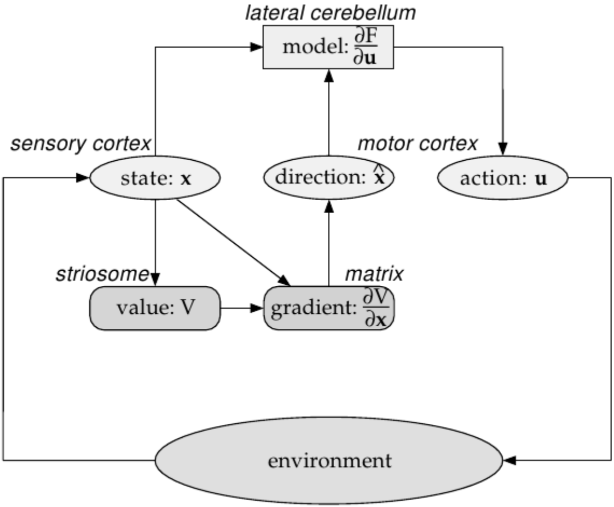
Model-free, stochastic action selection (Cortex and basal ganglia)

Unsupervised, supervised and reinforcement learning



Action selection with a forward model.

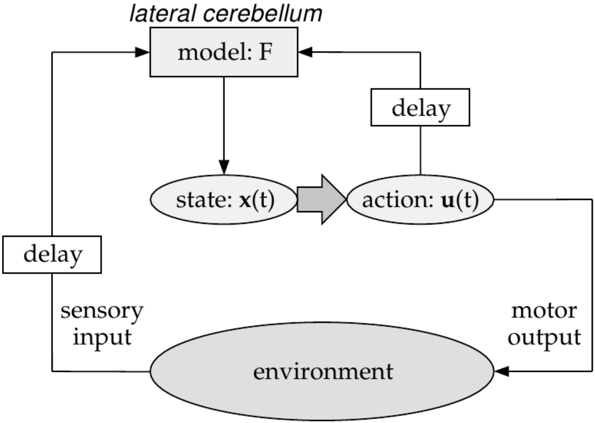
Unsupervised, supervised and reinforcement learning


lateral cerebellum
model: $\frac{\partial F}{\partial \mathbf{u}}$
sensory cortex
motor cortex
state: $\mathbf{x}$
direction: $\hat{\mathbf{x}}$
action: $\mathbf{u}$
striosome
matrix
value: V
gradient: $\frac{\partial V}{\partial \mathbf{x}}$
environment

Differential model-based action selection.

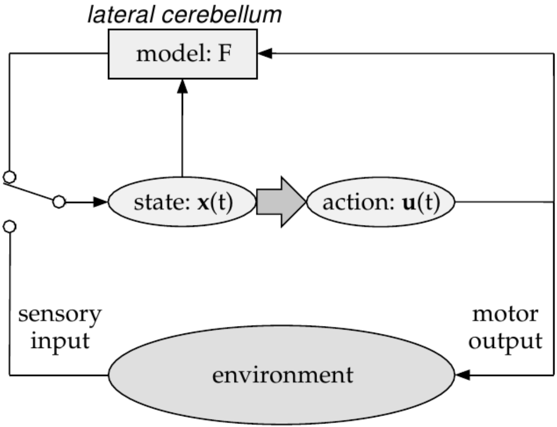
Doya, Kenji. Neural networks 12.7 (1999): 961-974.

Function of Doya's brain model



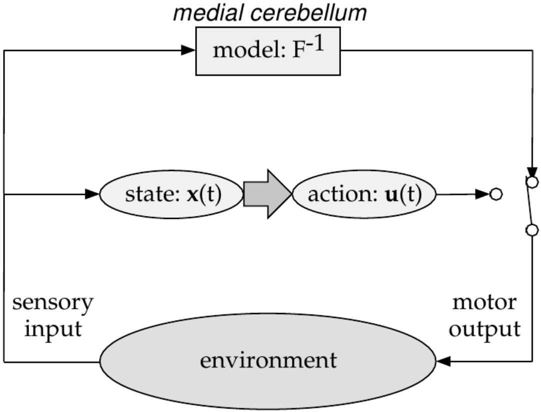
Compensation of sensory feedback delay with a forward model. The thick arrow represents either of the state-to-action mapping by the architectures shown above.

Function of Doya's brain model



Simulation of behaviour using a forward model instead of the real environment.

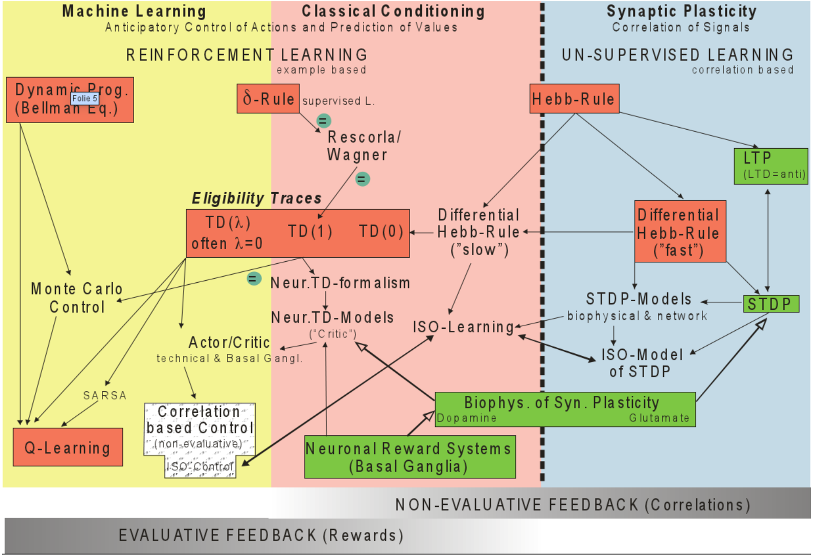
Function of Doya's brain model



Encapsulation of complex decision process in a simple reactive mapping.

Doya, Kenji. Neural networks 12.7 (1999): 961-974.

Learning (according to F. Wörgötter)



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)

Neural Modelling

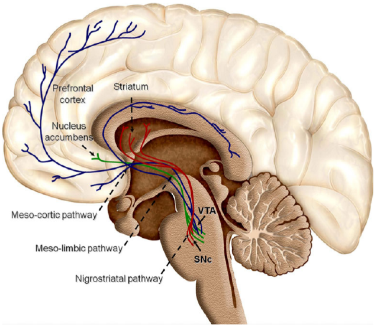
Reinforcement learning has revolutionised our understanding of learning in the brain in the last 20 years (Y. Niv)

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.

Suri, R. E., Schultz, W. (2001) Temporal difference model reproduces anticipatory neural activity. *Neural Computation* **13(4)**, 841-862.

Reward system in the brain (beyond algorithms)

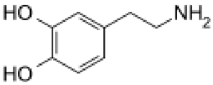


Overview of reward structures in the human brain. Dopaminergic neurons are located in the midbrain structures substantia nigra (SNc) and the ventral tegmental area (VTA).

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.

O. Arias-Carrión et al. "Dopaminergic reward system ..." Int. Archives of Medicine 3 (2010): 24.

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.

O. Arias-Carrión, E. Pöppel. Act Neurobiol Exp 67 (2007) 481-488

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.

K. Berridge, T. Robinson. Brain Res Brain Res Rev 28 (1998) 309-69.

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.

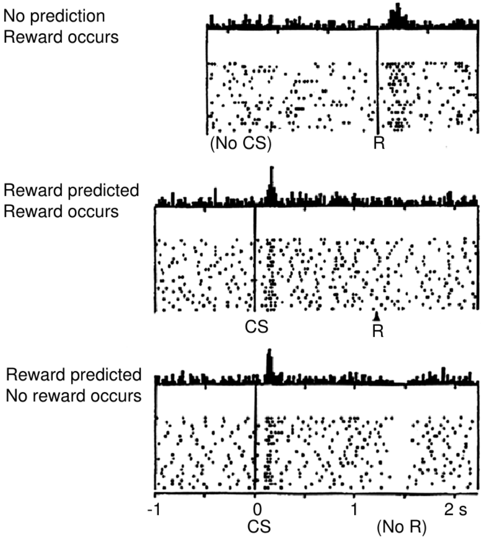
S. Peciña et al. J Neurosci 23 (2003) 9395-402.

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

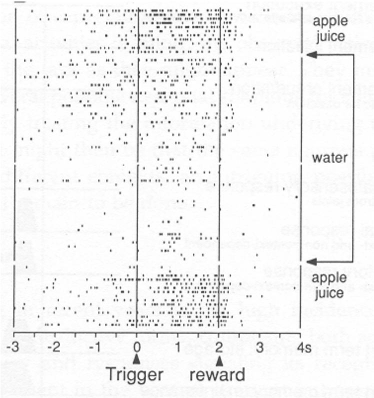
Dopamine neurons in ventral tegmental area (VTA)



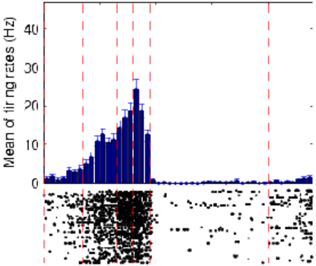
W. Schultz et al. Science 1997.

Striatum and learnt values

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



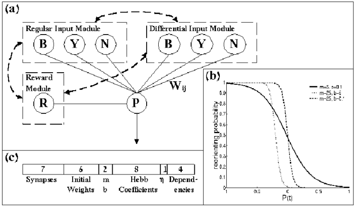
Schultz



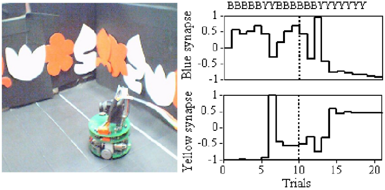
Daw: Ramping from start to food reward

P. Dayan

Evolution of reinforcement learning in a model



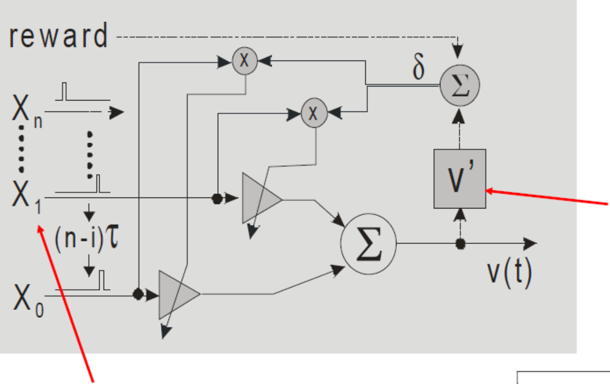
(a) The bee's neural network controller. (b) The bee's action function. Probability of reorienting direction of flight as a function of $P(t)$ for different values of parameters m; b. (c) The genome sequence of the simulated bee.



(a) The foraging robot. (b) Blue and yellow differential weights represent the expected rewards from the two flower colours along the trials. Top: Flower col or chosen in each trial. (blue flowers: 1/2 μl nectar, yellow: 1 μl in half the flowers, contingencies switched after trial 10.)

Y. Niv, D. Joel, I. Meilijson, E. Ruppin (2001) Evolution of reinforcement learning in uncertain environments: Emergence of risk aversion and matching. *Proc. ECAL*.

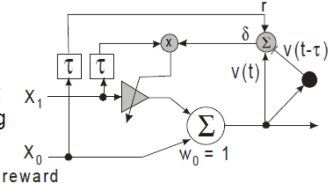
How to implement TD in a neuronal way (F. Wörgotter)



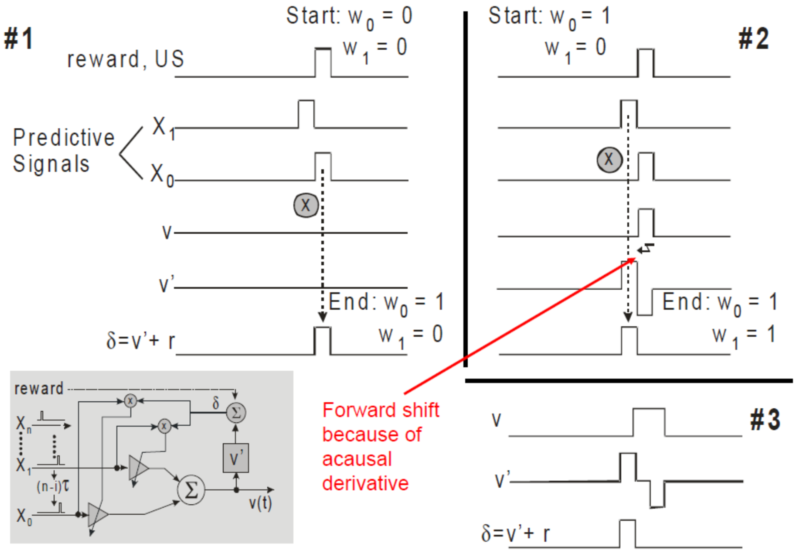
v(t+1)-v(t)

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.



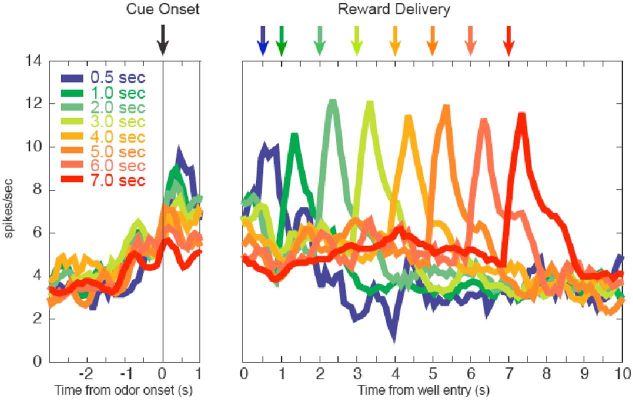
How to implement TD in a neuronal way (F. Wörgotter)



Forward shift because of acausal derivative

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?



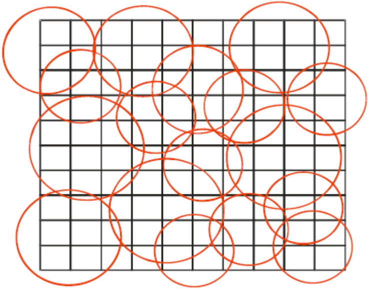
Average firing rate of 19 dopaminergic neurons, recorded in rats performing an odor-discrimination task in which one of the odors predicted that a reward would be delivered in a food-well, with some delay.

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.



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

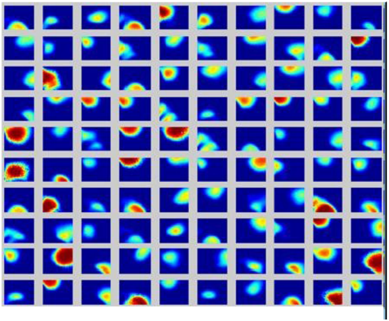
Goal: A simulated rat should find a reward in an arena.

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.

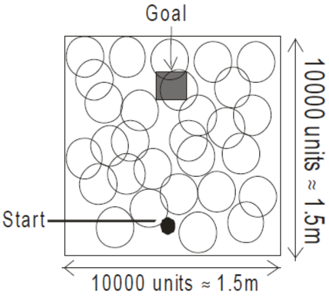
2) Rats (probably) use their hippocampal Place-Fields to learn such task. These place fields have different sizes and cover the space in an overlapping way. Furthermore, they fire to some degree stochastically.

   Hence they represent an Action Value Function Approximation system.

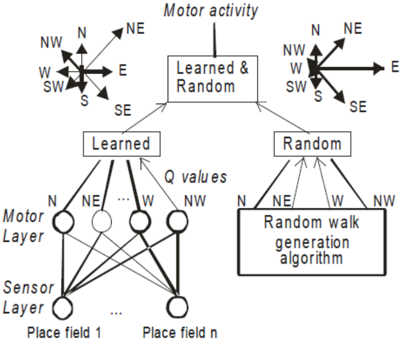


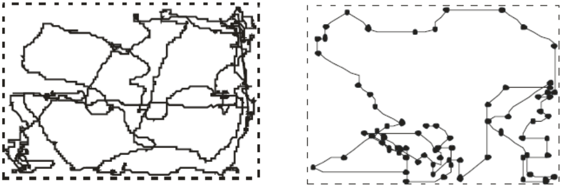
F. Wörgotter

Place field system



Path generation and Learning

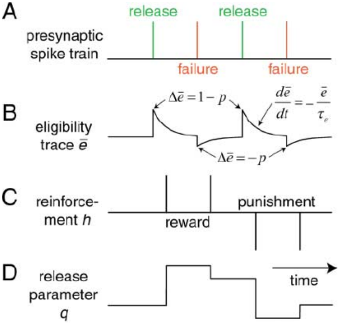




Real (left) and generated (right) path examples.

F. Wörgotter

Reinforcement learning in spiking neural networks

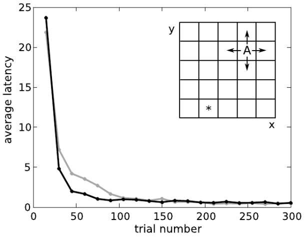


- Randomness of synaptic transmission is harnessed by the brain for learning
- Possible if synapses are "hedonistic", responding to a global reward signal by increasing their probabilities of vesicle release or failure, depending on which action immediately preceded re- ward.
- Hedonistic synapses learn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

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron, 40(6), 1063-1073.

A spiking neural network model of actor-critic learning

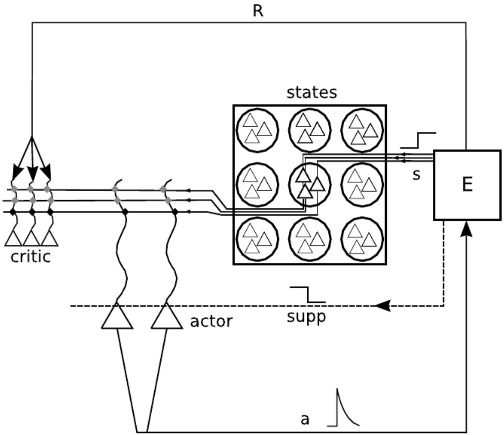
- Spiking neural network model implements actor-critic temporal-difference learning
- Combines local plasticity rules with a global reward signal
- Network solves a gridworld task with sparse rewards
- Similar equilibrium performance as look-up table variants



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

Potjans, W., Morrison, A., & Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. Neural computation, 21(2), 301-339.

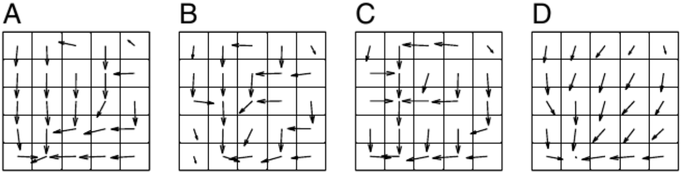
A spiking neural network model of actor-critic learning



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.

A spiking neural network model of actor-critic learning



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 is consistent with the temporal sensitivity displayed by neural prediction error signals?
  - Nature of (subjective) time and temporal order?
- How is unlearning represented?
  - Extinction learning: Pawlov's dog eventually stopped to drool

A few publications on neuro-biological RL

Y. Niv. Reinforcement learning in the brain. J. Math. Psychol. 53 (2009) 139-154.

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

W. Schultz (1998) Predictive reward signal of dopamine neurons. *J. Nphys.* **80**, 1-27.

Daw, N. D., and Touretzky, D. S. (2000) Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, **32-33**, 679-684 and (2001) Operant behavior suggests attentional gating of dopamine system inputs. *Neurocomputing* **38-40**, 1161-1167.

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

P Dayan & Y Niv (2008) - Reinforcement learning and the brain: The Good, The Bad and The Ugly - Current Opinion in Neurobiology, 18(2), 185-196

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron, 40(6), 1063-1073.

Doya, K. (2000). Reinforcement learning in continuous time and space. Neural computation, 12(1), 219-245.

Potjans, W., Morrison, A., & Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. Neural computation, 21(2), 301-339.

Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., & Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. PLoS computational biology, 5(12), e1000586.