Announcements:

Exam: Monday July 2, 4pm

Exam (counts 67%)

- paper and pencil
- no textbook, no slides, no calculator, no notes.
- similar to exercises and quizzes
- total approx. 45 points
  - 15 points for quiz-like questions (30 questions)
  - 30 points for exercise-like calculations
    - (4 exercises with a, b, c, d …)
- about 15 points are 'easy', 20 medium, 10 'difficult'

Miniprojects (count 33%).

- miniproject validated after 'fraud detection interview'

Recommended exam preparation

(1) do (or redo) exercises yourself
(2) if stuck, read the relevant chapter of the textbook
(see page 2 of slides)
(3) check the solution
(4) look at the quiz question (always orange slides)
(5) if stuck, read the relevant chapter of the textbook
(see page 2 of slides)

(the slides are useful only if you have annotated them yourself during the lecture; not a stand-alone tool)

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Objectives for today:

- **three-factor learning rules can be implemented by the brain**
- **eligibility traces as 'candidate parameter updates'**
- **the dopamine signal has signature of the TD error**
- **model-based versus model-free learning in the brain**

Today we finish at around 12:30

Reading for this week:

Sutton and Barto, Reinforcement Learning
(MIT Press, 2nd edition 2018, also online)

Chapter: 15

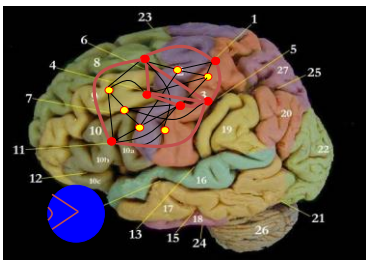
Background reading:

(1) Fremaux, Sprekeler, Gerstner (2013) Reinforcement learning using a continuous-time actor-critic framework with spiking neurons *PLOS Computational Biol. doi:10.1371/journal.pcbi.1003024*

(2) Fremaux, Gerstner (2016) Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules *Frontiers in neural circuits 9, 85*

(3) *J Gläscher, N Daw, P Dayan, JP O'Doherty (2010)* States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning, *Neuron 66 (4), 585-595*

Review: Artificial Neural Networks for action learning



Where is the supervisor?
Where is the labeled data?

Replaced by:
'Value of action'
- 'goodie' for dog
- 'success'
- 'compliment'

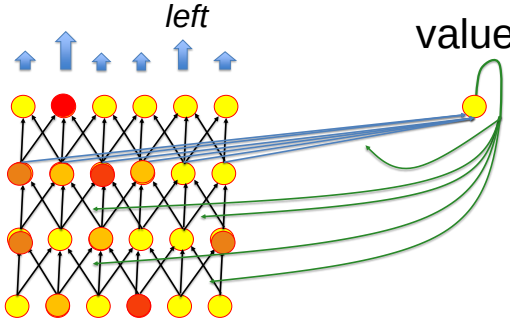
BUT:
Reward is rare:
'sparse feedback' after a long action sequence

Review: Actor-Critic = 'REINFORCE' with TD signal

- Estimate $V(s)$
- learn via TD error

actions

advance *push left*



value

$V(s)$

TD-error

$\delta = \eta[r_t + \gamma V(s') - V(s)]$

Review: Actor-Critic = 'REINFORCE' with TD signal

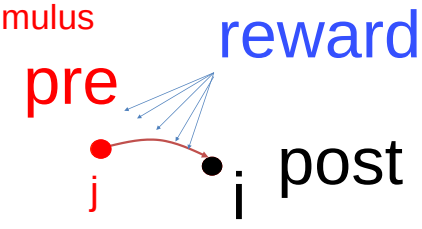
Actor–Critic with Eligibility Traces (episodic), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: trace-decay rates $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\lambda^{\mathbf{w}} \in [0,1]$; step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize S (first state of episode)
  $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ (d'-component eligibility trace vector)
  $\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ (d-component eligibility trace vector)
  $I \leftarrow 1$
  Loop while S is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{z}^{\mathbf{w}} \leftarrow \gamma\lambda^{\mathbf{w}}\mathbf{z}^{\mathbf{w}} + I\nabla\hat{v}(S,\mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \gamma\lambda^{\boldsymbol{\theta}}\mathbf{z}^{\boldsymbol{\theta}} + I\nabla \ln\pi(A|S,\boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\delta\mathbf{z}^{\mathbf{w}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\delta\mathbf{z}^{\boldsymbol{\theta}}$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

Review: Policy Gradient: Comparison with Biology

parameter = weight w_j

$$\Delta w_j \propto R(y,\vec{x})[\,y - \langle y\rangle]x_j$$



Weight vector turns in direction of input

Three factors: **reward** **post** **pre**

$$\Delta w_{ij} = \eta\, R(\vec{y},\vec{x})[\,y_i - \langle y_i\rangle]x_j$$

postsynaptic factor is

'activity – expected activity'

Questions for today:

- does the brain implement reinforcement learning algorithms?
- Can the brain implement an actor-critic structure?

Previous slide.

Reinforcement Learning includes a set of very powerful algorithm – as we have seen in previous lecture.

For today the big question is:
Is the structure of the brain suited to implement reinforcement learning algorithms?

If so which one? Q-learning or SARSA?

Is the brain architecture compatible with an actor-critic structure?

These are the questions we will address in the following.
And to do so, we have to first get a big of background information on brain anatomy.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy

Previous slide.

Before we can make a link to Reinforcement Learning we need to know a bit more about the brain.

1. Coarse Brain Anatomy and Reinforcement Learning

Reinforcement learning needs:

- states / sensory representation
- action selection
- reward signals

Previous slide.

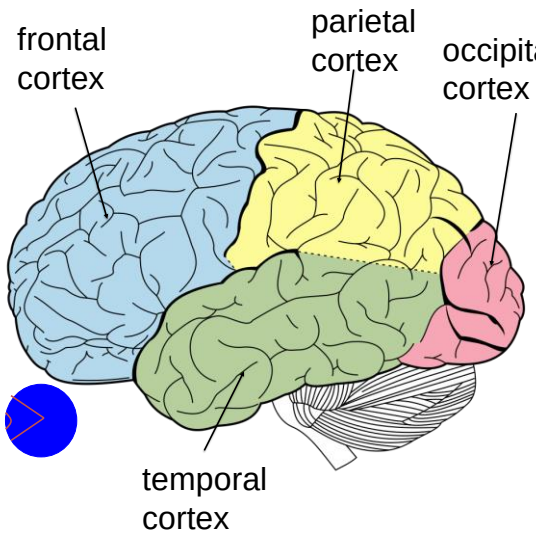
In reinforcement learning, the essential variables are the states (defined by sensory representation), a policy for action selection, the actions themselves, and the rewards given by the environment.

If we want to link reinforcement learning to the brain, we will have to search for corresponding substrates and functions in the brain.
Therefore we now take a rather coarse and simplified look at the anatomy of the brain.

The Wikipedia articles give more information for those who are interested.

1. Coarse Brain Anatomy: Cortex

Sensory representation in visual/somatosensory/auditory cortex



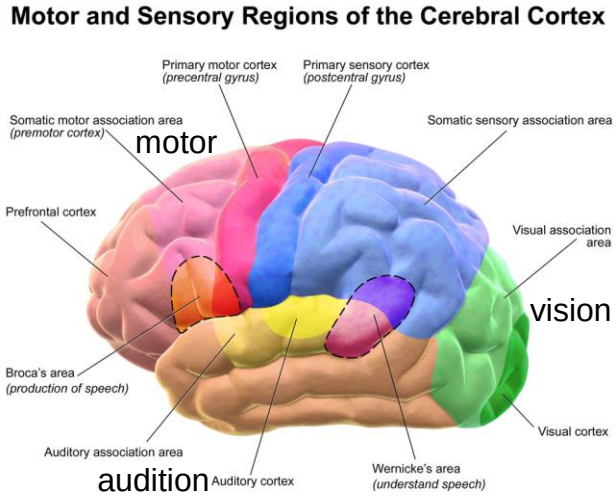


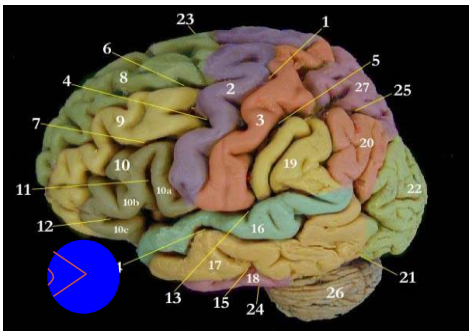
fig: Wikipedia

Previous slide.
Left: Anatomy. The Cortex is the part of the brain directly below the skull. It is a folded sheet of densely packed neurons. The biggest folds separate the four main part of cortex (frontal, Parietal, occipital, and temporal cortex)

Right: Functional assignments. Different parts of the brain are involved in different tasks. For example there several areas involved in processing visual stimuli (called primary and secondary cortex). Other areas are involved in audition (auditory cortex) or the presentation of the body surface (somatosensory cortex). Yet other areas are prepared in the preparation of motor commands for e.g., arm movement.

1. Coarse Brain Anatomy

- many different cortical areas
- but also several brain nuclei sitting below the cortex
- Some of these nuclei send dopamine signals
- Dopamine is related to **reward**, surprise, and pleasure
- Dopamine from: VTA, nucleus accumbens, substantia nigra



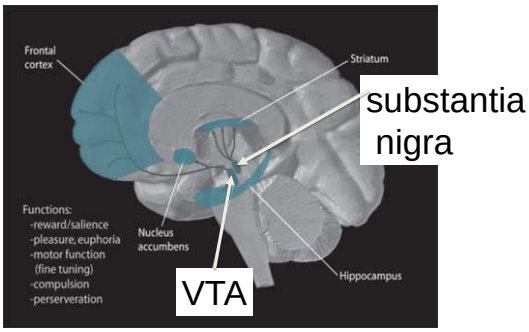


fig: Wikipedia commons

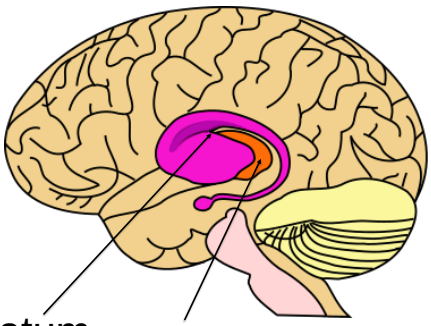
Previous slide.
Left: Anatomy. View on the folds of the cortex, and main cortical areas in different color.

Right: Below the cortex sit different nuclei. Some of these nuclei use dopamine as their signaling molecule. Important nuclei for dopamine are the Ventral Tegmental Area (VTA) and the Substantia Nigra pars compacte (SNc). These dopamine neurons send their signals to large areas of the cortex.
Since dopamine is involved in reward, these dopamine neurons will play a role in this lecture that links reinforcement learning and the brain.

1. Coarse Brain Anatomy: Striatum

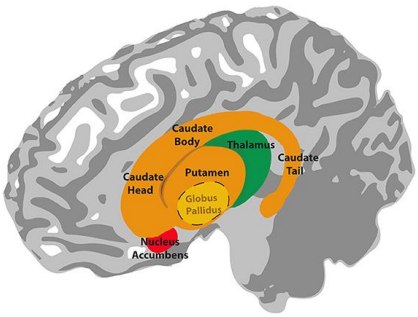
- Striatum sits below cortex
- Part of the 'basal ganglia'
- **Dorsal striatum** involved in **action selection**, decisions

Striatum consists of
- Caudate
- Putamen



striatum

thalamus



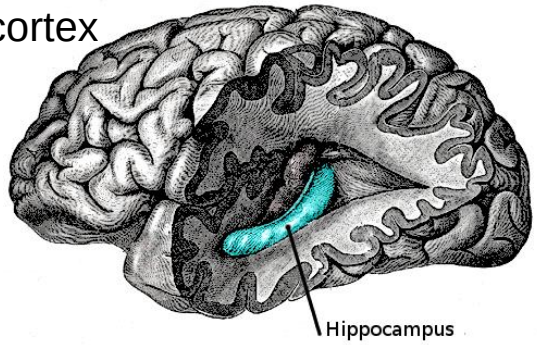
Nucleus Accumbens is part of **ventral striatum**

fig: Wikipedia

Previous slide.
Left: Sketch of the Anatomical location of striatum and thalamus.
Right: the striatum lies also below the cortex. Since the striatum is involved in action selection it will play an important role in this lecture.
From Wikipedia:
The **striatum** is a nucleus (a cluster of neurons) in the subcortical basal ganglia of the forebrain. The striatum is a critical component of the motor and reward systems; receives glutamatergic and dopaminergic inputs from different sources; and serves as the primary input to the rest of the basal ganglia.
Functionally, the striatum coordinates multiple aspects of cognition, including both motor and action planning, decision-making, motivation, reinforcement, and reward perception.The striatum is made up of the caudate nucleus and the lentiform nucleus. The lentiform nucleus is made up of the larger putamen, and the smaller globus pallidus.
In primates, the striatum is divided into a **ventral striatum**, and a **dorsal striatum**, subdivisions that are based upon function and connections. The ventral striatum consists of the nucleus accumbens and the olfactory tubercle. The dorsal striatum consists of the caudate nucleus and the putamen. A white matter, nerve tract (the internal capsule) in the dorsal striatum separates the caudate nucleus and the putamen.[4] Anatomically, the term *striatum* describes its striped (striated) appearance of grey-and-white matter

1. Coarse Brain Anatomy: hippocampus

Hippocampus
- Sits below/part of temporal cortex
- Involved in memory
- Involved in spatial memory



Spatial memory:
knowing where you are,
knowing how to navigate in an environment

fig: Wikipedia

Henry Gray (1918) *Anatomy of the Human Body*

Previous slide.
From Wikipedia:

The **hippocampus** (named after its resemblance to the seahorse, from the Greek ἱππόκαμπος, "seahorse" from ἵππος *hippos*, "horse" and κάμπος *kampos*, "sea monster") is a major component of the brains of humans and other vertebrates. Humans and other mammals have two hippocampuses, one in each side of the brain. The hippocampus belongs to the limbic system and plays important roles in the consolidation of information from short-term memory to long-term memory, and in spatial memory that enables navigation. The hippocampus is located under the cerebral cortex (allocortical)[1][2][3] and in primates in the medial temporal lobe.

1. Coarse Brain Anatomy and Reinforcement Learning

Reinforcement learning needs:

- states / sensory representation → cortex?, hippocampus?
- action selection → striatum?, motor cortex?
- reward signals → dopamine?

Previous slide.

In reinforcement learning, the essential variables are the states (defined by sensory representation), a policy for action selection, the actions themselves, and the rewards given by the environment.

If we want to link reinforcement learning to the brain, we will have to search for corresponding substrates and functions in the brain.

The above rough ideas need to be defined during the rest of this lecture.

1. Quiz: Coarse Functional Brain anatomy

[ ] the brain = the cortex (synonyms)
[ ] the cortex consists of several areas
[ ] some areas are more involved in vision, others more in the representation of the body surface
[ ] below the cortex there are groups (clusters) of neurons
[ ] Hippocampus sends out dopamine signals
[ ] VTA and nucleus accumbens send out dopamine signals
[ ] dopamine is linked the reward, pleasure, surprise
[ ] striatum is involved in action selection

Previous slide. Your comments

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy
2. Synaptic Plasticity

Previous slide.

Reinforcement Learning is, obviously, a form of ‘learning’. Learning is related to synaptic plasticity. Therefore this is our second topic.

2. Behavioral Learning

Learning actions:

→ riding a bicycle
→ play tennis
→ play the violin

Remembering episodes

→ first day at EPFL
→ plan how to get home

Build ‘models of the world’

Previous slide.

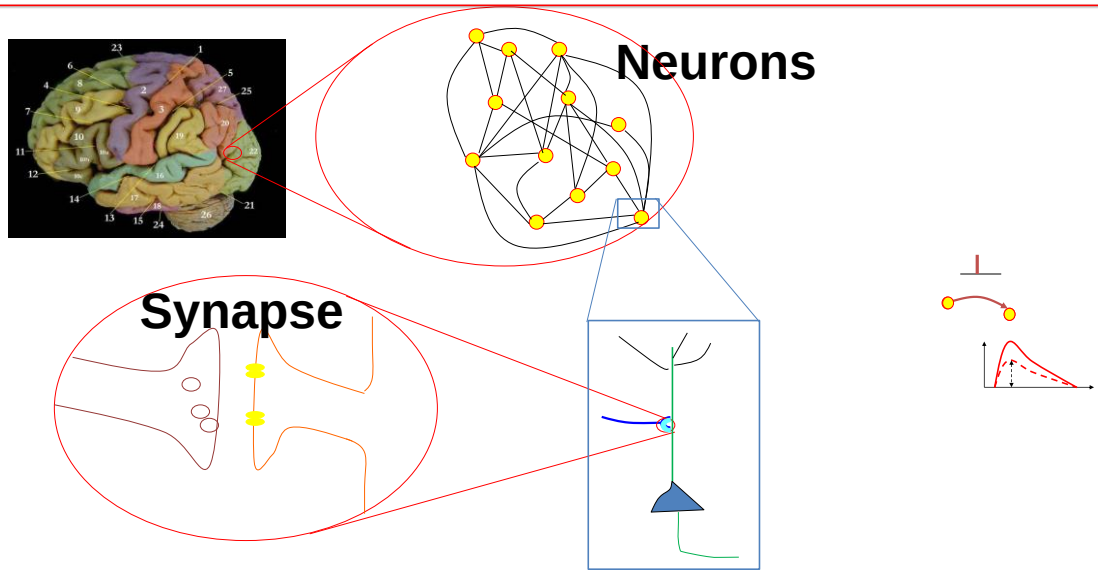
When we learn to ride a bike we learn with Reinforcement-like feedback, e.g., we don’t want to fall because falling hurts.

When we learn play the tennis or the violin we also get feedback via the observed outcome – which can be good or bad.

When we walk around a city for the first time we develop a model of the environment – even in the absence of any specific rewards (except, may be, that it is good to know how to find the way home).

All these are examples of learning. The last one might be unsupervised learning, but the others are clearly reinforcement learning.

2. Behavioral Learning – and synaptic plasticity



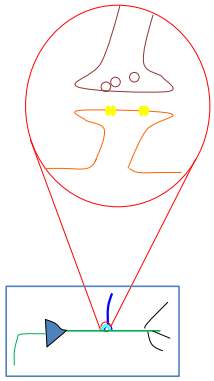
Synaptic Plasticity =Change in Connection Strength

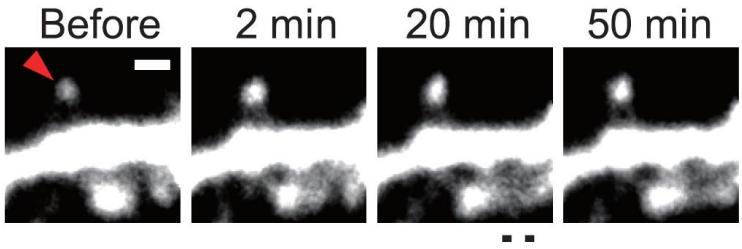
Previous slide.

When we observe learning on the level of behavior (we get better at tennis), then this implies that something has changed in our brain:
The contact points between neurons (called synapses) have changed. Synaptic changes manifest themselves as a change in connections strength.

Synaptic plasticity describes the phenomena and rules of synaptic changes.

2. Synaptic plasticity – structural changes





Yagishita et al.
Science, 2014

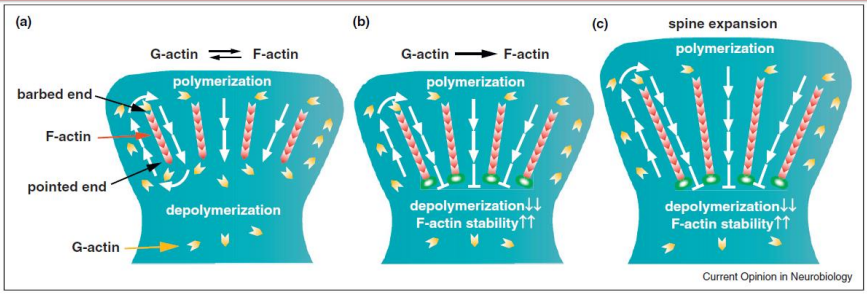
Previous slide.

The synaptic connection consists of two parts. The end of an axonal branch coming from the sending neuron; and the counterpart, a protrusion on the dendrite of the receiving neuron, called spine.

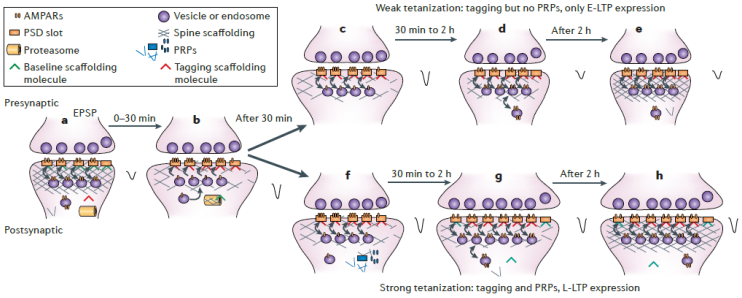
We refer to the sending neuron as presynaptic and to the receiving one as postsynaptic.

A change in the connection strength is observable with imaging methods as an increase in the size of the spine. The bigger spine remains big for a long time (here observed for nearly one hour).

2. synaptic plasticity – molecular changes



Bosch et al. 2012, Curr. Opinion Neurobiol.

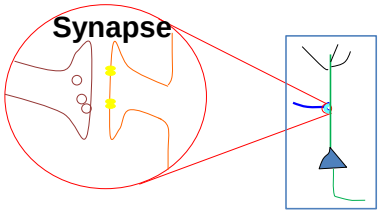


Redondo and Morris 2011, Nature Rev. Neurosci.

Previous slide.

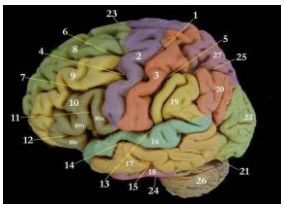
The actual molecular machinery inside the spine is very complicated – and of no further interest of us in the following.

2. synaptic plasticity – connections change



More space for fingers allocated in cortex
- musicians vs. non-musicians

Amunts et al. Human Brain Map. 1997
Gaser and Schlaug, J. Neuosci. 2003



More space allocated in hippocampus
- London taxi driver vs bus driver

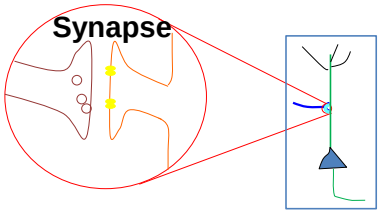
Macquire et al. Hippocampus 2006

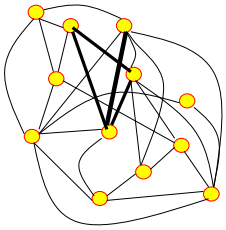
Previous slide.

We said at the beginning of the lecture that different areas of the brain are involved in different tasks. For example, the somatosensory cortex represents the body surface. Nowadays one can measure that the size of the cortical area devoted to fingers is larger for musicians than for non-musicians. Since musicians are not born with a larger area, this result implies that synaptic plasticity can influence the function of the neurons in the brain.

Similarly, hippocampus is involved in spatial navigation. Not surprisingly, London taxi drivers have a bigger hippocampus than London bus drivers.

2. Synaptic plasticity: summary





- Connections can be strong or weak
- Strong connections have thick spines
- Synaptic plasticity
  = change of connection

Syn. Plasticity should enable Learning

- adapt to the statistics of task and environments (useful filters, allocate space etc)
- memorize facts and episodes
- learn models of the world
- learn motor tasks

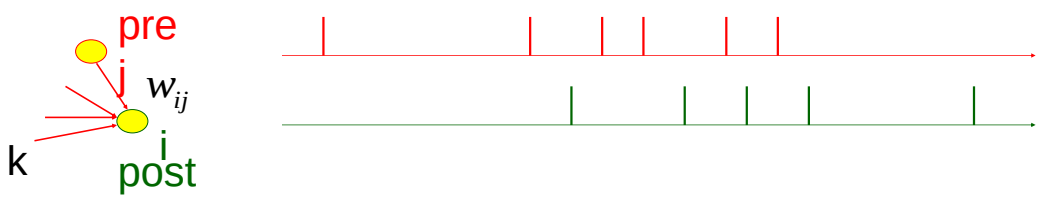
Previous slide.

Thus connections can be strong or weak – and synaptic plasticity describes the changes of one synapse from weak to strong or back.

The synaptic changes are thought to be the basis of learning – whatever the learning task at hand.

The question now is: Are the any rules that would predict whether and when a synapse gets stronger?

2. Hebb rule / Hebbian Learning



When an axon of cell **j** repeatedly or persistently takes part in firing cell **i**, then j's efficiency as one of the cells firing i is increased

Hebb, 1949

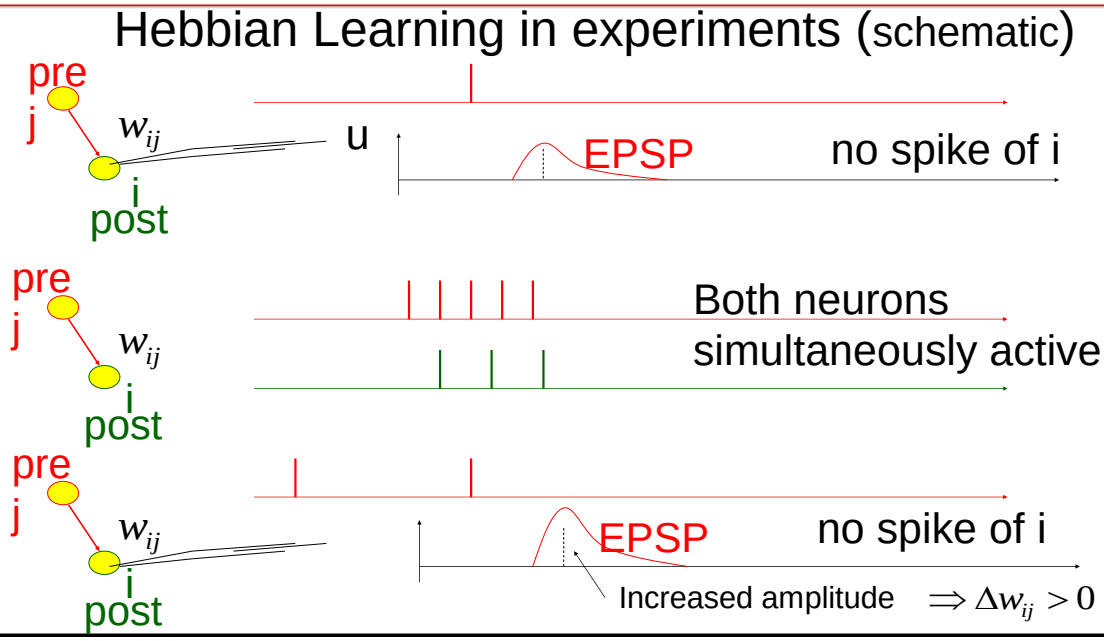
- local rule
- simultaneously active (correlations)

Previous slide.

The Hebb rule is the classic rule of synaptic plasticity.

It is often summarized by saying: if two neurons are active together, the connection between those two neurons gets stronger.
Note that the original formulation of Hebb also has a 'causal' notion: 'takes part in firing' – which is more than just firing together.

2. Synaptic plasticity: Long-Term Potentiation (LTP)



Previous slide.

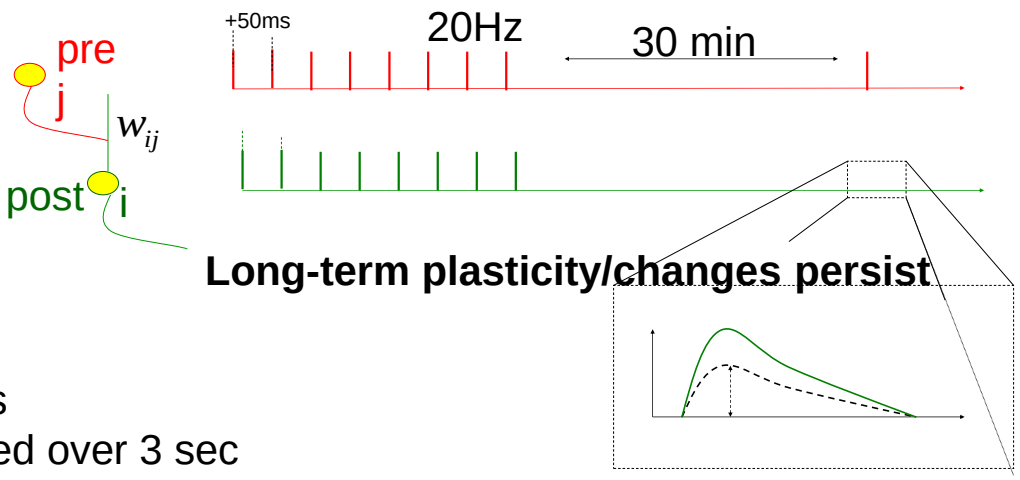
In a schematic experiment,

1) You first test the size of the synapse by sending a pulse from the presynaptic neurons across the synapses. The amplitude of the excitatory postsynaptic potential (EPSP) is a convenient measure of the synaptic strength. It has been shown that it is correlated with the size of the spine.

2) Then you do the Hebbian protocol: you make both neurons fire together

3) Finally you test again the size of the synapse. If the amplitude is bigger you conclude that the synaptic weight has increased.

2. Why the name ‘ Long-term plasticity ‘ (LTP)?



Changes
- induced over 3 sec
- persist over 1 – 10 hours (or longer?)

Previous slide.

Experimentalists talk about Long-Term Potentiation (LTP), because once the change is induced it persists for a long time. Interestingly, it is sufficient to make the two neurons fire together for just a few seconds.

Thus inducation of plasticity is rapid, but the changes persist for an hour or more.

2. Classical paradigm of LTP induction – pairing

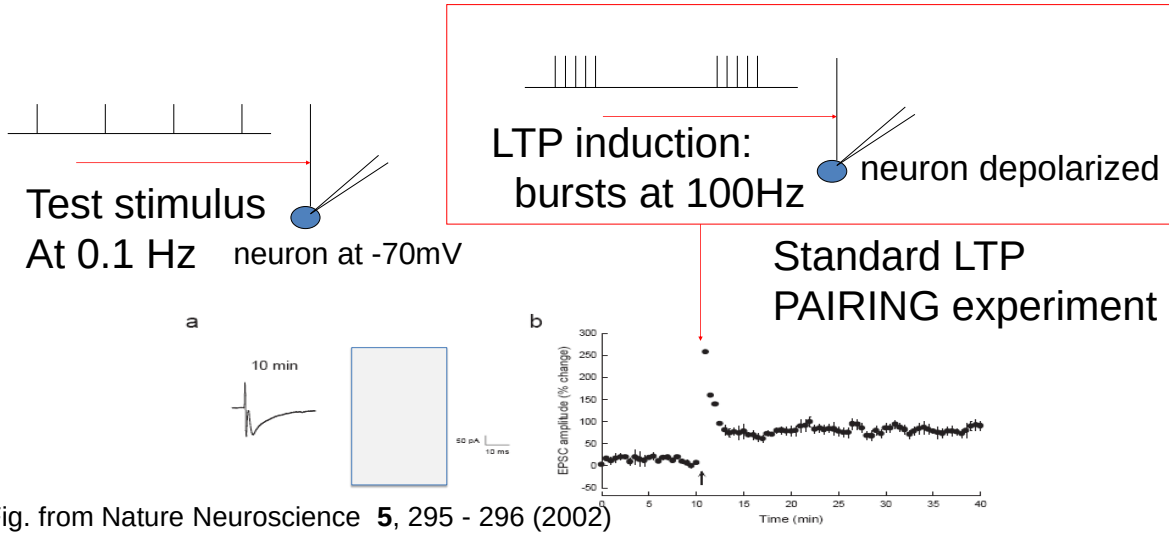
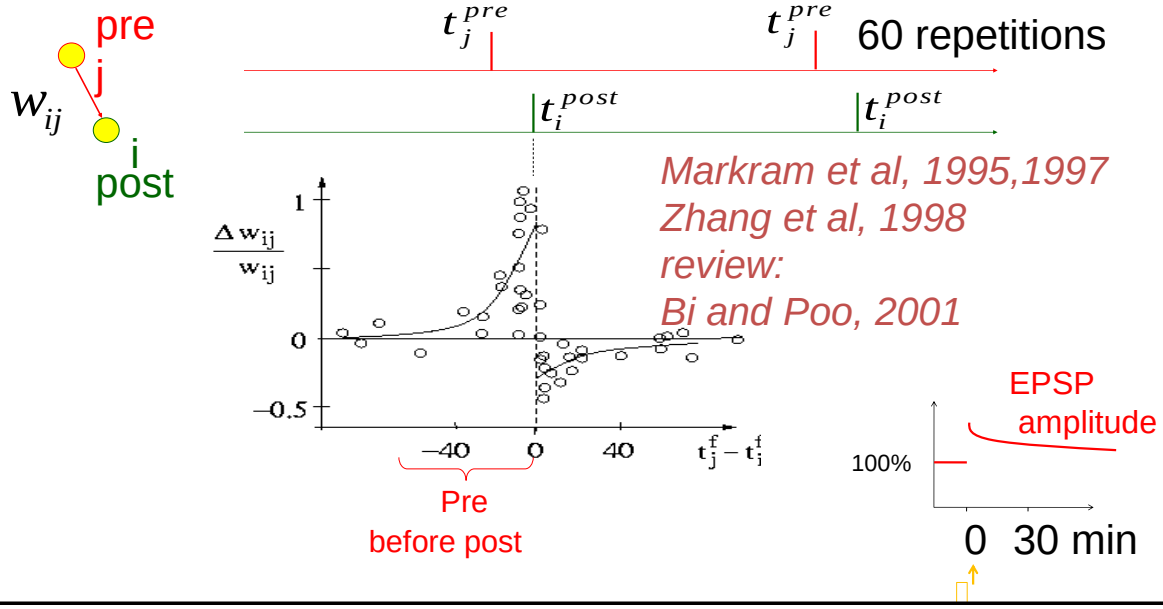


Fig. from Nature Neuroscience **5**, 295 - 296 (2002)
D. S.F. Ling, … & Todd C. Sacktor
See also: Bliss and Lomo (1973), Artola, Brocher, Singer (1990), Bliss and Collingridge (1993)

Previous slide.
In one classic paradigm of LTP induction, the presynaptic fibers are strongly stimulated (with bursts of 100 pulses per second, repeated several times) while the postsynaptic neuron is stimulated with an electrode to put above its normal 'resting potential'.

The size of the synapses is measured by the excitatory postsynaptic current (EPSC) which is itself proportional to the EPSP. After the stimulation (which lasts less than a minute) the synapse remains strong for a long time.
The initial transient is of no importance for our discussion.

2. Spike-timing dependent plasticity (STDP)



Previous slide.
In second paradigm of LTP induction, the presynaptic neuron is stimulated so that it emits a single spike, and the postsynaptic neuron is also stimulated so that emits a single spike – either a few milliseconds before or after the presynaptic spike. This stimulation protocol (for example pre-before-post) is then repeated several times.

The increase of the synaptic weight (induced by repeated pre-before-post) persists for a long time.

Since the size of the increase depends on the relative timing of the two spikes, this induction protocol is called Spike-Timing-Dependent Plasticity (STDP).

2. Summary: Synaptic plasticity

Synaptic plasticity
- makes connections stronger or weaker
- can be experimentally induced
- needs 'joint activation' of the two connected neurons
- is induced rapidly, but can last for a long time
- Spike-timing dependent plasticity is one of many protocols

Hebb rule: 'neurons that fire together, wire together'

S. Loewl and W. Singer, Science 1992

Previous slide.

There are several experimental paradigms to induce synaptic changes.
Most of these paradigms are consistent with the Hebb rule:
Neurons that fire together, wire together, a slogan that was introduced by Loewl and Singer in 1992.

However, in all these Hebbian learning rules and their corresponding experimental paradigms, the role of reward is unclear and not considered.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

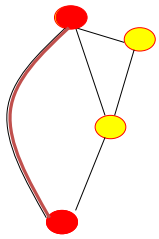
1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules

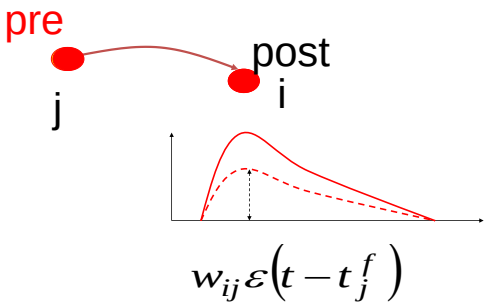
Previous slide.

Since Hebbian learning rules are limited, we have to extend the framework and include a ‘third factor’ that could represent reward.

3. Classification of synaptic changes: unsupervised learning

Hebbian Learning
= unsupervised learning





$$w_{ij}\,\varepsilon\left(t - t_j^f\right)$$

$$\Delta w_{ij} \propto F(pre, post)$$

Previous slide.

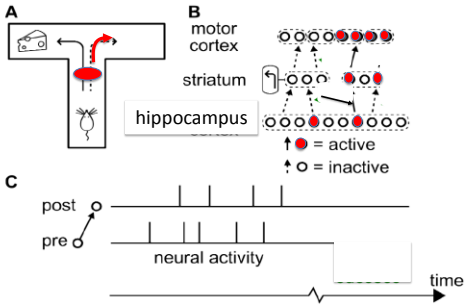
In standard Hebbian learning, the change of the synaptic weight depends only on presynaptic activity and the state of the postsynaptic neuron. The rule is local, and does not contain the notion of reward or success.

3.Limits of unsupervised learning

Is Hebbian Learning sufficient?
No!

Image: Gerstner et al. NEURONAL DYNAMICS,

Eligibility trace:
Synapse keeps memory of pre-post Hebbian events



Dopamine: Reward/success

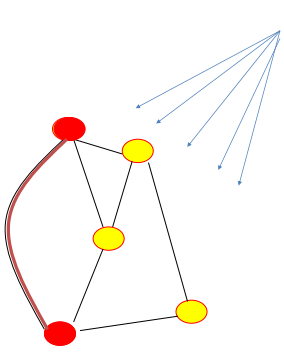
Schultz et al. 1997; Waelti et al., 2001;

→ Reinforcement learning: success = reward – (expected reward)

TD-learning, SARSA, Policy gradient (book: Sutton and Barto, 1997/2018)

Previous slide.
Hebbian learning as it stands is not sufficient to describe learning in a setting were rewards play a role. If joint activity of pre- and post causes stronger synapses, the rat is likely to repeat the same unrewarded action a second time.
Hypothetical functional role of neuromodulated synaptic plasticity. Reward-modulated learning
(A) Schematic reward-based learning experiment.Ananimal learns to perform a desired sequence of actions(e.g.,move straight,then turn left) in a T-maze through trial-and-error with rewards (cheese)
(B) The current position (“place”) of the animal in the environment is represented by an assembly of active cells in the hippocampus.These cells feed neurons (e.g.,in the dorsal striatum) which code for high-level actions at the choice point,e.g., “turn left” or “turn right.” These neuronsin turn project to motorcortex neurons,responsible for the detailed implementation of actions. A success signal modulates (green arrows) the induction of plasticity
(C) Neuromodulator timing. While spikes occur on the time scale of milliseconds, the success signal may come a few seconds laters.

3. Classification of synaptic changes: Reinforcement Learning



Reinforcement Learning = reward + Hebb

$$\Delta w_{ij} \propto F(pre, post, SUCCESS)$$

local global

broadly diffused signal: neuromodulator

Previous slide.

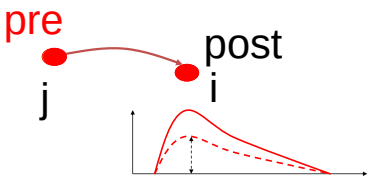
For the moment we say the reinforcement learning depends on three factors: the Hebbian pre- and postsynaptic factor plus a success signal related to reward.
We will get more precise later.

3. Classification of synaptic changes

unsupervised vs reinforcement

LTP/LTD/Hebb
Theoretical concept

- passive changes
- exploit statistical correlations

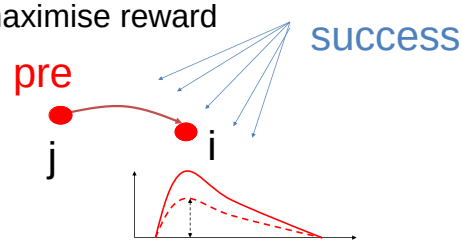


Functionality

-useful for development
(develop good filters)

Reinforcement Learning
Theoretical concept

- conditioned changes
- maximise reward



Functionality

- useful for learning
a new behavior

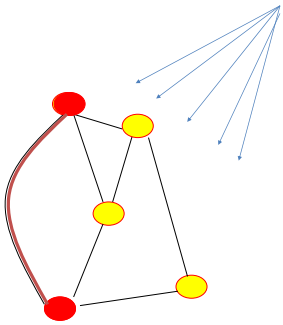
Previous slide.

This does not mean the standard Hebbian learning is wrong: in fact it is very useful for the development of generic synaptic connections, e.g., to make neurons develop good filtering properties that pick up relevant statistical signals in the stream of input.

The three-factor rules are relevant for learning novel behaviors via feedback through reward.

3. Three-factor rule of Hebbian Learning

= Hebb-rule gated by a neuromodulator



Neuromodulators: Interestingness, surprise; attention; novelty

$$\Delta w_{ij} \propto F(pre, post, MOD)$$

local global

Previous slide.

The three-factor rules have a Hebbian component: pre- and postsynaptic activity together, but in addition the third factor which is related to neuromodulators.

There are several neuromodulators in the brain

Neuromodulator projections

- 4 or 5 neuromodulators
- near-global action

Dopamine/reward/TD:
Schultz et al., 1997,
Schultz, 2002

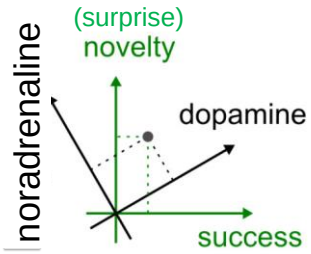
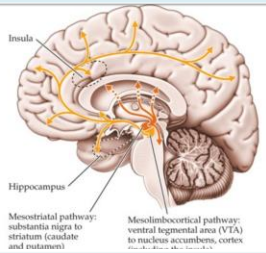


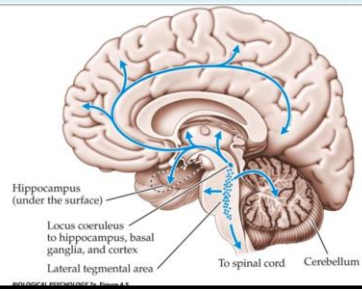
Image:
Fremaux and Gerstner, Frontiers (2016)

Image: Biological Psychology, Sinauer

Dopamine (DA)



Noradrenaline (NE)



Previous slide.

The most famous neuromodulator is dopamine (DA) which is related to reward, as we will see.

But there are other neuromodulators such as noradrenaline (also called norepinephrine, NE) which is related to surprise.

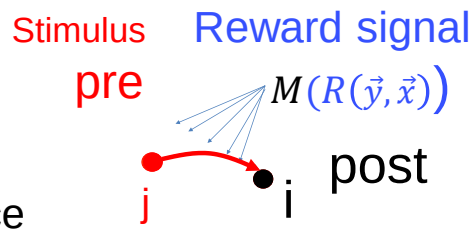
Left: the mapping between neuromodulators and functions is not one-to-one. Indeed, dopamine also has a 'surprise' component.

Right: most neuromodulators send axons to large areas of the brain, in particular to several cortical areas. The axons branch out in thousands of branches. Thus the information transmitted by a neuromodulator arrives nearly everywhere. In this sense, it is a 'global' signal, available in nearly all brain areas.

3. Formalism of Three-factor rules

x_j = activity of presynaptic neuron

y_i = activity of presynaptic neuron



Step 1: co-activation sets eligibility trace

$$\Delta z_{ij} = \eta \; f(y_i)\, g(x_j)$$

Step 2: eligibility trace decays over time

$$z_{ij} \leftarrow \lambda \, z_{ij}$$

Step 3: eligibility trace translated into weight change

$$\Delta w_{ij} = \eta \, M\left(R(\vec{y},\vec{x})\right) z_{ij}$$

Previous slide.

Three-factor rules are implementable with eligibility traces.

The joint activation of pre- and postsynaptic neuron sets a ‘flag’. This step is similar to the Hebb-rule, but the change of the synapse is not yet implemented.

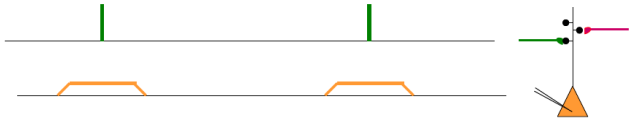
The eligibility trace decays over time

However, if a neuromodulatory signal M arrives before the eligibility trace has decayed to zero, an actual change of the weight is implemented.

3. Hebbian LTP versus Three-factor rules

Hebbian coactivation: (i)
pre-post-post-post

Hebbian coactivation: (ii)
but no post-spikes



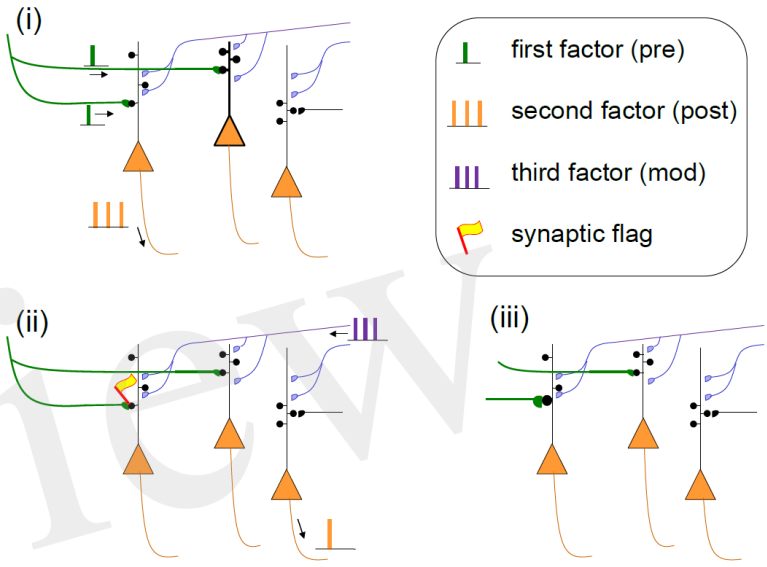
Scenario of three-factor rule: Hebb+modulator (iii)

Previous slide.

The joint activation of pre- and postsynaptic neuron sets a 'flag'. This step is similar to the Hebb-rule, but the change of the synapse is not yet implemented. Note that joint activation can imply spikes of pre- (green) and postsynaptic (orange) neuron (top);
Or spikes of a presynaptic neuron combined with a weak voltage increase in the postsynaptic neuron (middle).

Bottom: three-factor rule only if a neuromodulatory signal M arrives before the eligibility trace has decayed to zero, an actual change of the weight is implemented. The neuromodulater arrives through the branches

3. Three-factor rules: synaptic flags and delayed reward (mod)



synaptic flag plays role of eligibility trace

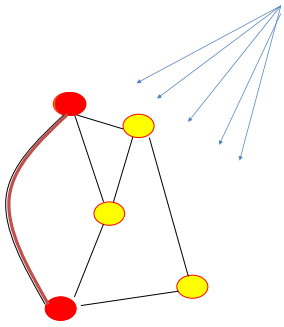
Fig: Gerstner et al. 2018

Previous slide.

Specificity of three-factor learning rules.
(i) Presynaptic input spikes (green) arrive at two different neurons, but only one of these also shows postsynaptic activity (orange spikes).
(ii) A synaptic flag is set only at the synapse with a Hebbian co-activation of pre- and postsynaptic factors; the synapse become then eligible to interact with the third factor (blue). Spontaneous spikes of other neurons do not interfere.
(iii) The interaction of the synaptic flag (eligibility trace) with the third factor leads to a strengthening of the synapse (green).

Fig caption: Gerstner et al. 2018

3. Recent experiments for Three-factor rules



Neuromodulators for reward, interestingness, surprise; attention; novelty

Step 1: co-activation sets eligibility trace

Step 2: eligibility trace decays over time

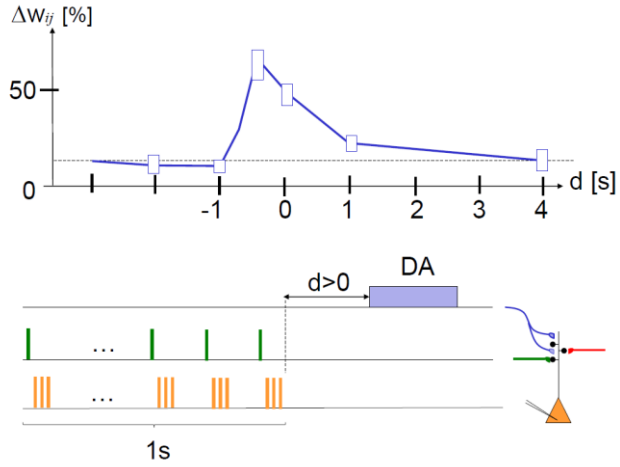
Step 3: delayed neuro-Modulator:
eligibility trace translated into weight change

Previous slide.

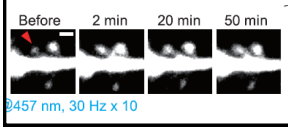
three-factor learning rules are a theoretical concept.

But are there any experiments? Only quite recently, a few experimental results were published that directly address this question.

3. Three-factor rules in striatum: eligibility trace and delayed Da



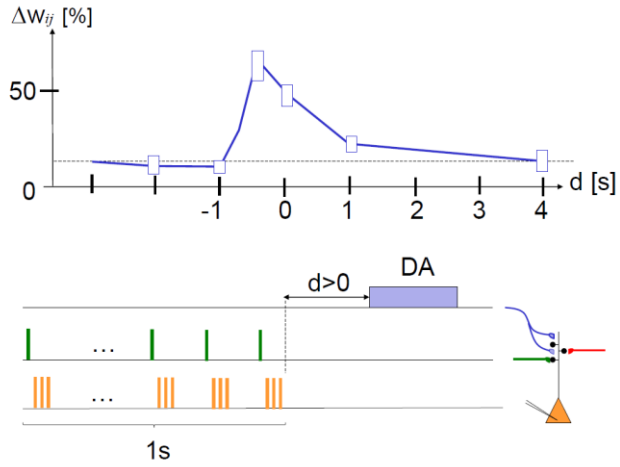
Yagishita et al. 2014



2457 nm, 30 Hz x 10

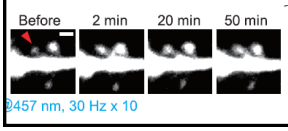
-Dopamine can come with a delay of 1s
-Long-Term stability over at least 50 min.

3. Three-factor rules in striatum: eligibility trace and delayed Da



Yagishita et al. 2014

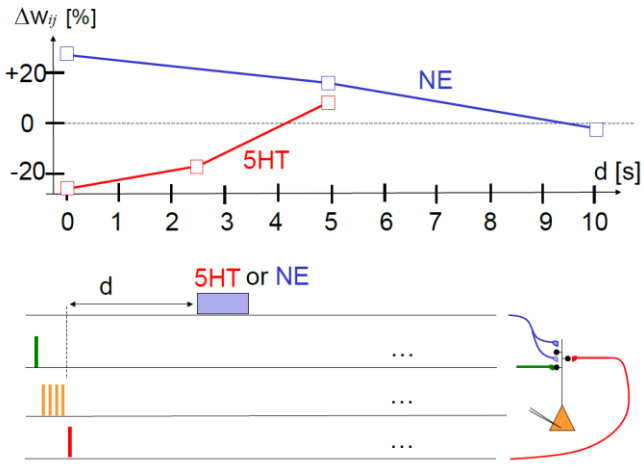
In striatum medial spiny cells, stimulation of presynaptic glutamatergic fibers (green) followed by three postsynaptic action potentials (STDP with pre-post-post-post at +10ms) repeated 10 times at 10Hz yields LTP if dopamine (DA) fibers are stimulated during the presentation (d < 0) or shortly afterward (d = 0s or d = 1s) but not if dopamine is given with a delay d = 4s; redrawn after Fig. 1 of (Yagishita et al., 2014), with delay d defined as time since end of STDP protocol



2457 nm, 30 Hz x 10

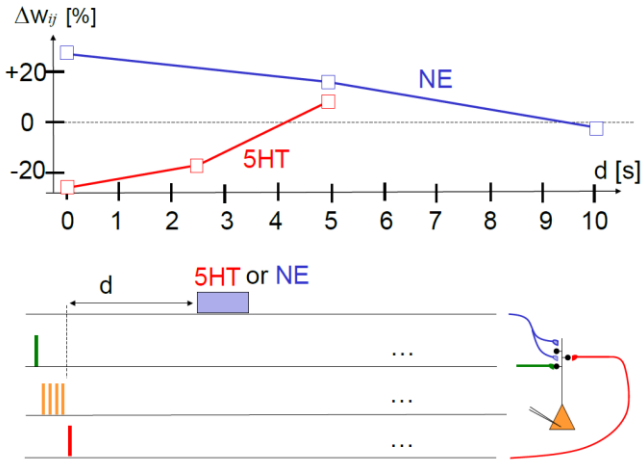
-Dopamine can come with a delay of 1s
-Long-Term stability over at least 50 min.

3. Three-factor rules in cortex: eligibility trace and delayed NE



(He et al., 2015).

3. Three-factor rules in cortex: eligibility trace and delayed NE



In cortical pyramidal cells, stimulation of two independent presynaptic pathways (green and red) from layer 4 to layer 2/3 by a single pulse is paired with a burst of four postsynaptic spikes (orange).

If the pre-before-post stimulation was combined with a pulse of norepinephrine (NE) receptor agonist isoproterenol with a delay of 0 or 5s, the protocol gave LTP (blue trace).

If the post-before-pre stimulation was combined with a pulse of serotonin (5-HT) of a delay of 0 or 2.5s, the protocol gave LTD (red trace).

(He et al., 2015).

3. Three-factor rules: summary

Three factors are needed for synaptic changes:

- Presynaptic factor = spikes of presynaptic neuron
- Postsynaptic factor = spikes of postsynaptic neuron or increased voltage
- Third factor = Neuromodulator such as dopamine

Previous slide.

three-factor learning rules are a theoretical concept.

But recent experiments show that the brain really can implement three-factor rules. Importantly, the third factor (neuromodulator) can come with a delay of one or two seconds after the Hebbian induction protocol that sets the eligibility trace.

Quiz. Synaptic Plasticity and Learning Rules

Standard Long-term potentiation

[ ] has an acronym LTP
[ ] takes more than 10 minutes to induce
[ ] lasts more than 30 minutes
[ ] depends on presynaptic activity
AND on state of postsynaptic neuron

Learning rules

[ ] Hebbian learning depends on presynaptic activity
AND on state of postsynaptic neuron
[ ] Reinforcement learning depends on neuromodulators
such as dopamine indicating reward
[ ] Three-factor rule: presynaptic signal, postsynaptic signal, and neuromodulator signal (e.g., DA) must arrive at the same time.

Previous slide.

Your comments.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited

Previous slide.

I now want to show that reinforcement learning with policy gradient gives rise to three-factor learning rules.

4. Review from week 10 .Policy Gradient over multiple time steps

Calculation yields several terms of the form

Total accumulated discounted reward collected in one episode starting at s_t, a_t

$$\Delta\theta_j \propto [R^{a_t}_{s_t \to s_{end}}] \frac{d}{d\theta_j} \ln[\pi(a_t|s_t,\theta)]$$

$$+\gamma[R^{a_{t+1}}_{s_{t+1} \to s_{end}}] \frac{d}{d\theta_j} \ln[\pi(a_{t+1}|s_{t+1},\theta)]$$

$$+ \ldots$$

Previous slide.
This is a repetition from week 12.

4. Policy Gradient over multiple time steps (Exercise last week)

Step 1: Rewrite

Step 2: Use same update formula, but for state s_{t+1}

Step 3: Reorder terms according to r_{t+n}

$$\Delta\theta_j \propto [R^{a_t}_{s_t \to s_{end}}] \frac{d}{d\theta_j} \ln[\pi(a_t|s_t,\theta)]$$
$$+\gamma[R^{a_{t+1}}_{s_{t+1} \to s_{end}}] \frac{d}{d\theta_j} \ln[\pi(a_{t+1}|s_{t+1},\theta)]$$
$$+ \ldots$$

Previous slide.
This is a repetition of the exercises from week 10.

4. Policy Gradient for eligibility traces (Exercise last week)

Step 4: Introduce 'shadow variables' for eligibility trace

$z_k \leftarrow z_k \ \lambda$ decay of **all** traces

$z_k \leftarrow z_k + \frac{d}{dw_k}\ln[\pi(a|s, w_k)]$ increase of **all** traces

Step 5: Rewrite update rule for parameters with eligibility trace

$\Delta w_k = \eta \ r_t \ z_k$

Previous slide.
This is a repetition of the exercises from week 10.

4. Eligibility traces from Policy Gradient (Exercise last week)

Run trial. At each time step, observe state, action, reward

1) Update eligibility trace

$z_k \leftarrow z_k \ \lambda$ decay of **all** traces

$z_k \leftarrow z_k + \frac{d}{dw_k}\ln[\pi(a|s, w_k)]$ increase of **all** traces

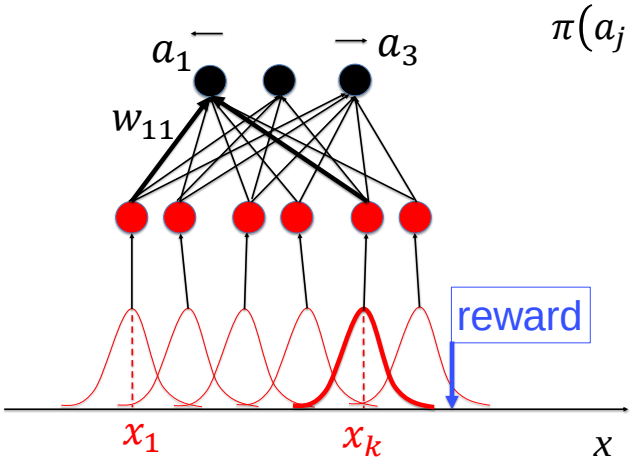
2) update parameters

$\Delta w_k = \eta \ r_t \ z_k$

Previous slide.
This is a repetition of the exercises from week 10.

4. Example: Linear activation model with softmax policy

left: $a_1=1$ *stay:* $a_2=1$ *right:* $a_3=1$



parameters

$$\pi(a_j = 1 | x, \theta) = softmax[\sum_k w_{jk}\, y_k]$$

$$y_k = f(x - x_k)$$

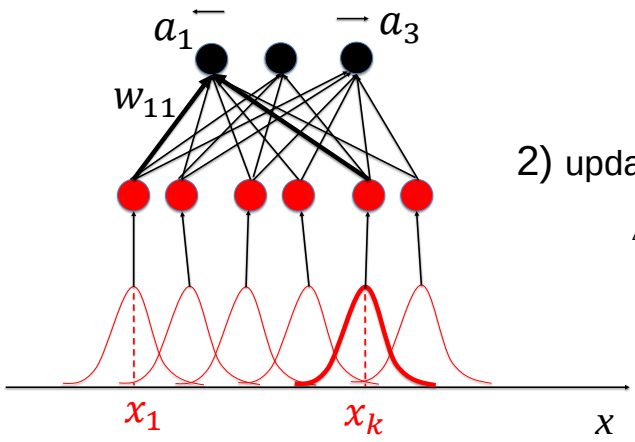
f=basis function

Previous slide.
Suppose the agent moves on a linear track.
There are three possible actions: left, right, or stay.

The policy is given by the softmax function. The total drive of the action neurons is a linear function of the activity y of the hidden neurons which in turn depends on the input x. The activity of hidden neuron k is f(x-x_k). The basis function f could for example be a Gaussian function with center at x_k.

4. Example: Linear activation model with softmax policy

left: *stay:* *right:*



1) Update eligibility trace (for each weight)

$$z_{ik} \leftarrow z_{ik}\,\lambda$$

$$z_{ik} \leftarrow z_{ik} + \frac{d}{dw_k}\ln[\pi(a_i|x)]$$

2) update weights

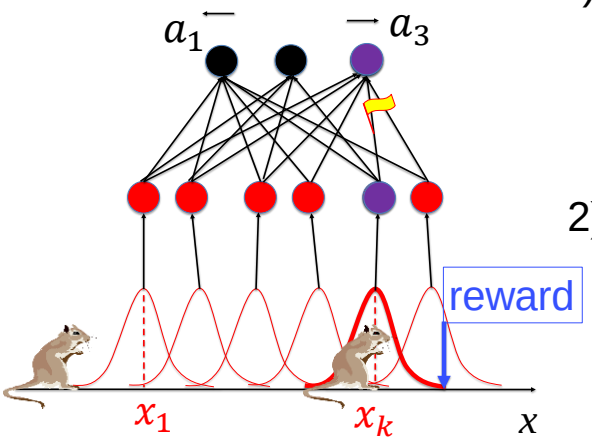
$$\Delta w_{lk} = \eta\; r_t\; z_{lk}$$

Exercise 1 now
8 minutes

Previous slide.
Now we apply the update rule resulting from policy gradient with eligibility traces descent (copy from earlier slide).
This is the in-class exercise (Exercise 1 of this week).

4. Example: Linear activation model with softmax policy

left: $a_1=1$ *stay:* $a_2=1$ *right:* $a_3=1$



0) Choose action $a_i \in \{0,1\}$

1) Update eligibility trace

$$z_{ik} \leftarrow z_{ik}\,\lambda$$

$$z_{ik} \leftarrow z_{ik} + y_k(x)[a_i - \pi(a_i|x)]$$

2) update weights

$$\Delta w_{lk} = \eta\ r_t\ z_{lk}$$

Previous slide.
This is the result of the in-class exercise (Exercise 1 of this week).
Importantly, the update of the eligibility trace is a local learning rule that depends on a presynaptic factor and a postsynaptic factor.

4. Summary: 3-factor rules from Policy

- Policy gradient with one hidden layer and linear softmax readout yields a 3-factor rule
- Eligibility trace is set by joint activity of presynaptic and postsynaptic neuron
- Update happens proportional to reward and eligibility trace
- The presynaptic neuron represents the state
- The postsynaptic neuron the action
- True online rule
  → could be implemented in biology
  → can also be implemented in parallel asynchr. hardware

Previous slide.
Summary: A policy gradient algorithm in a network where the output layer has a linear drive with softmax output leads to a three-factor learning rule for the connections between neurons in the hidden layer and the output.

These three factor learning rules are important because they are completely asynchronous, local, and online and could therefore be implemented in biology or parallel hardware.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited
5. Third factor

Previous slide.
So far the third factor remained rather abstract. We mentioned that a neuromodulator such as dopamine could be involved. Let us make this idea more precise and show experimental data.

5. Neuromodulators as Third factor

Three factors are needed for synaptic changes:

- Presynaptic factor = spikes of presynaptic neuron
- Postsynaptic factor = spikes of postsynaptic neuron or increased voltage
- Third factor = Neuromodulator such as dopamine

Presynaptic and postsynaptic factor 'select' the synapse.
→ a small subset of synapses becomes 'eligible' for change.
The 'Third factor' is a nearly global signal
→ broadcast signal, potentially received by all synapses.
Synapses need all three factors for change

Previous slide.
Before we start let us review the basics of a three-factor learning rule. We said that the third factor could be a neuromodulator such as dopamine.

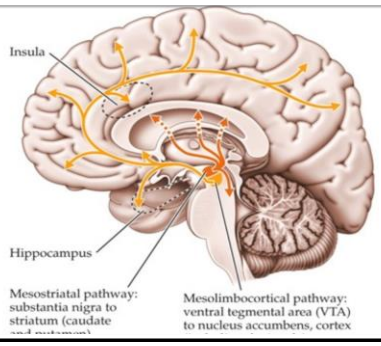
Review from week 8: Reward information

Neuromodulator **dopamine**: - is nearly globally broadcasted
- signals reward minus expected reward

‘success signal’

Schultz et al., 1997,
Waelti et al., 2001
Schultz, 2002

Dopamine



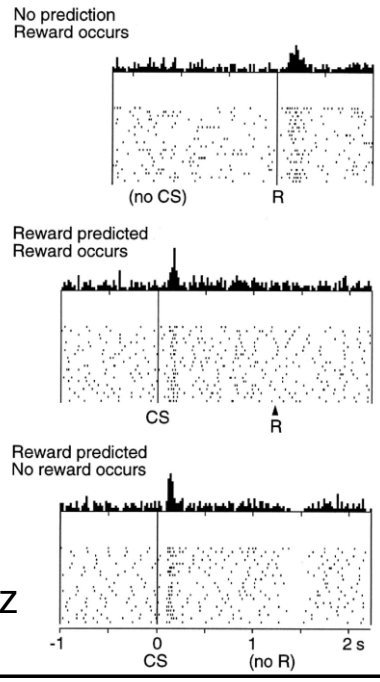
Previous slide. Dopamine neurons send dopamine signals to many neurons and synapses in parallel in a broadcast like fashion.

5. Dopamine as Third factor

Conditioning:
red light →1s→reward

CS:
Conditioning
Stimulus

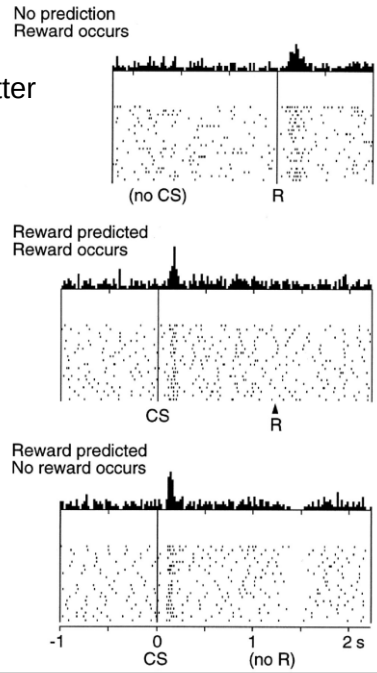
Sutton book, reprinted from W. Schultz



5. Dopamine as Third factor

This is now the famous experiment of W. Schultz.
In reality the CS was not a red light, but that does not matter

Figure 15.3: The response of dopamine neurons drops below baseline shortly after the time when an expected reward fails to occur. Top: dopamine neurons are activated by the unpredicted delivery of a drop of apple juice. Middle: dopamine neurons respond to a conditioned stimulus (CS) that predicts reward and do not respond to the reward itself. Bottom: when the reward predicted by the CS fails to occur, the activity of dopamine neurons drops below baseline shortly after the time the reward is expected to occur. At the top of each of these panels is shown the average number of action potentials produced by monitored dopamine neurons within small time intervals around the indicated times. The raster plots below show the activity patterns of the individual dopamine neurons that were monitored; each dot represents an action potential. From Schultz, Dayan, and Montague, A Neural Substrate of Prediction and Reward, *Science*, vol. 275, issue 5306, pages 1593-1598, March 14, 1997. Reprinted with permission from AAAS.



5. Summary: Dopamine as Third factor

- Dopamine signals ‘reward minus expected reward’
- Dopamine signals an ‘event that predicts a reward’
- Dopamine signals approximately the TD-error

$$DA(t) = [\underbrace{r(t) - (V(s) - \gamma V(s'))}_{\text{TD-delta}}]$$

Previous slide.

The paper of W. Schultz has related the dopamine signal to some basic aspects of Temporal difference Learning. The Dopamine signal is similar to the TD error.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

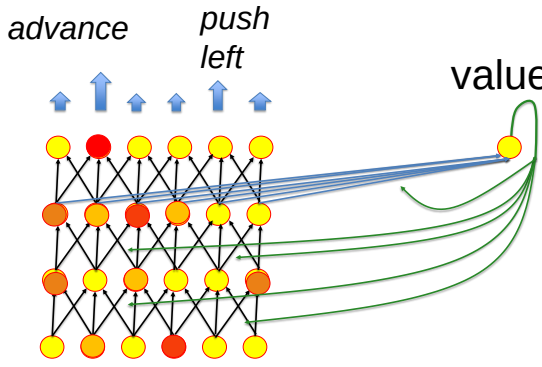
1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited
5. Third factor
6. Actor-critic revisited

Previous slide.

The next step is to put the three-factor learning rule in an actor-critic architecture

6. Review: Actor-Critic = 'REINFORCE' with TD signal

- Estimate $V(s)$
- learn via TD error



Previous slide.

Review of actor-critic architecture

6. Eligibility Traces with TD in Actor-Critic

Idea:

- keep memory of previous 'candidate updates'
- memory decays over time
- Update an **eligibility trace for each parameter**

$z_k \leftarrow z_k \ \lambda$ decay of **all** traces

$z_k \leftarrow z_k + \frac{d}{dw_k}\ln[\pi(a|s, w_k)]$ increase of **all** traces

- update **all** parameters:

$\Delta w_k = \eta \ [r - (V(s) - \gamma V(s'))] \ z_k$

TD-delta

→ policy gradient with eligibility trace and TD error

Previous slide.

Review of algorithm with actor-critic architecture and policy gradient with eligibility traces and TD.

6. Summary: Eligibility Traces with TD in Actor-Critic

Three-factor rules:

Presynaptic and postsynaptic factor 'select' the synapse.
→ a small subset of synapses becomes 'eligible' for change

The 'Third factor' is a nearly global broadcast signal
→ potentially received by all synapses.

Synapses need all three factors for change

The 'Third factor' can be the Dopamine-like TD signal
→ Need actor-critic architecture to calculate $\gamma V(s') - V(s)$
→ Dopamine signals $[r_t + \gamma V(s') - V(s)]$

Previous slide.

The three factor rule, dopamine, TD signals, value functions now all fit together.

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

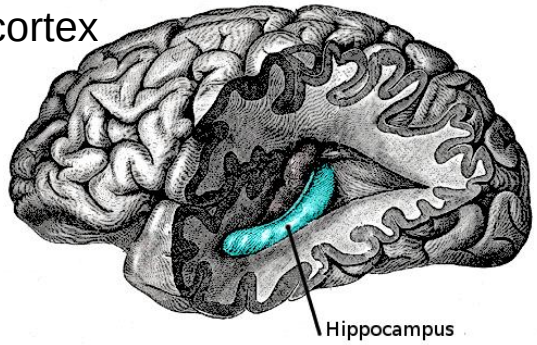
1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited
5. Third factor
6. Actor-critic revisited
7. Example of Navigation

Previous slide.

We said that the three factor rule, dopamine, TD signals, value functions now all fit together. Let's apply this to the problem of navigation.

7. Coarse Brain Anatomy: hippocampus

Hippocampus
- Sits below/part of temporal cortex
- Involved in memory
- Involved in spatial memory



Spatial memory:
knowing where you are,
knowing how to navigate in an environment

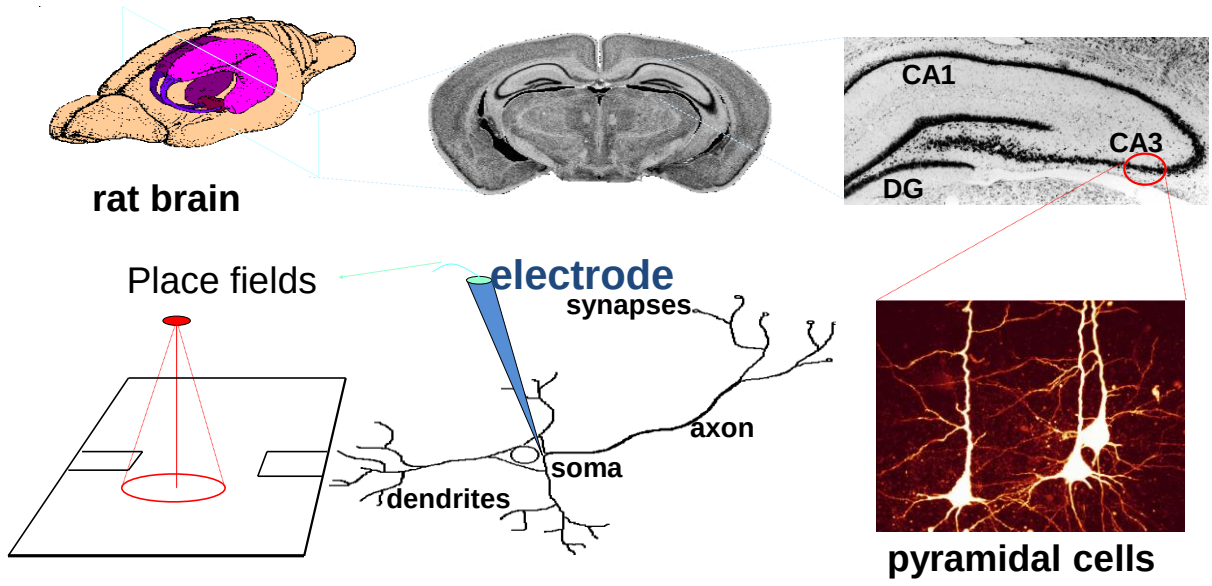
fig: Wikipedia

Henry Gray (1918) *Anatomy of the Human Body*

Previous slide.

the problem of navigation needs the spatical representation of the hippocampus.

7. Place cells in rat hippocampus

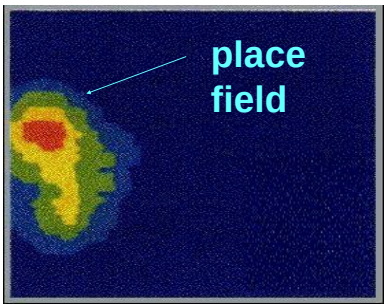


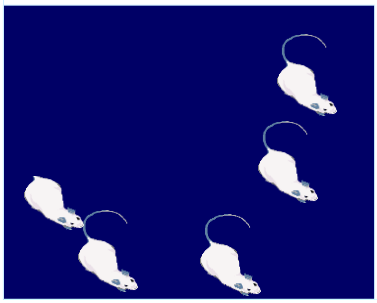
Previous slide.

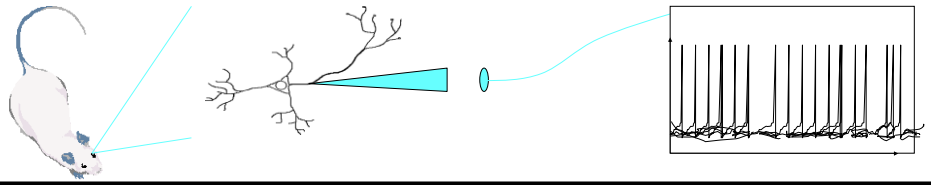
the hippocampus of rodents (rats or mice) looks somewhat different to that of humans. Importantly, cells in hippocampus of rodents respond only in a small region of the environment. For this reason they are called place cells. The small region is called the place field of the cell.

7. Hippocampal place cells

Main property: encoding the animal's location



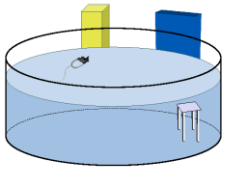


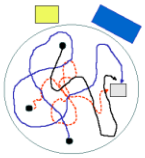


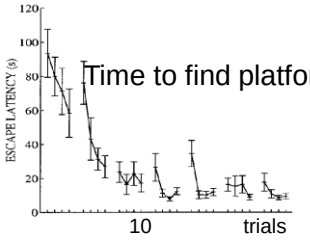
Previous slide.
Left: experimentally measured place field of a single cell in hippocampus.
Right: computer animation of place field

Review from week8: animal conditioning

Morris Water Maze







Time to find platform

Rats learn to find
the hidden platform

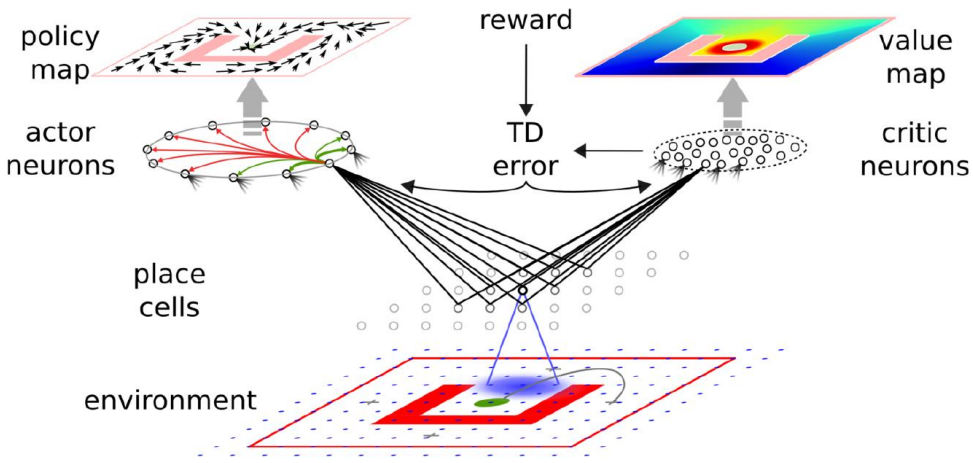
(Because they like to
get out of the cold water)

Foster, Morris, Dayan 2000

Previous slide.
Behvioral experiment in the Morris Water Maze.
The water is milky so that the platform is visible.

After a few trials the rat swims directly to the platform

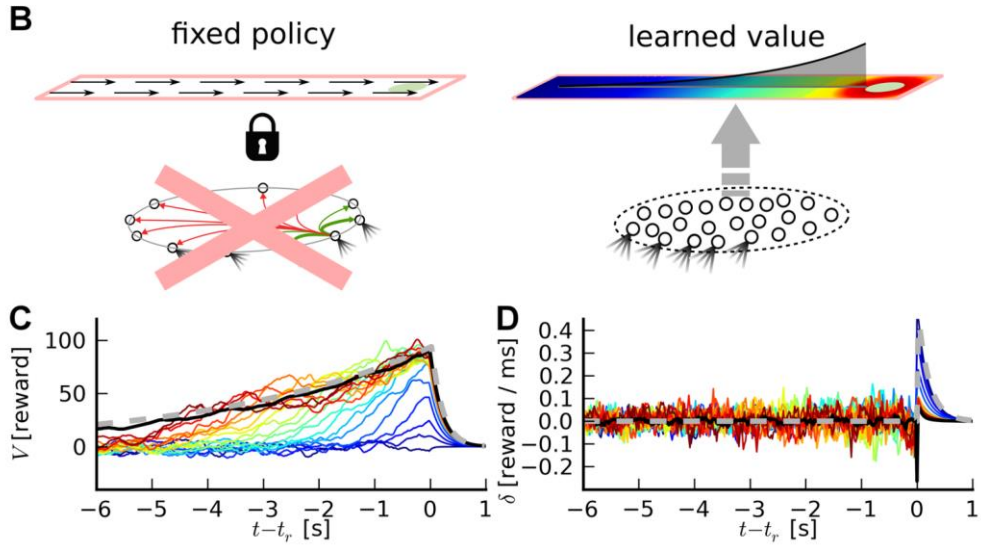
7. Maze Navigation with TD in Actor-Critic



Fremaux et al. (2013)

Figure 1. Navigation task and actor-critic network. From bottom to top: the simulated agent evolves in a maze environment, until it finds the reward area (green disk), avoiding obstacles (red). Place cells maintain a representation of the position of the agent through their tuning curves. Blue shadow: example tuning curve of one place cell (black); blue dots: tuning curves centers of other place cells. Right: a pool of critic neurons encode the expected future reward (value map, top right) at the agent's current position. The change in the predicted value is compared to the actual reward, leading to the temporal difference (TD) error. The TD error signal is broadcast to the synapses as part of the learning rule. Left: a ring of actor neurons with global inhibition and local excitation code for the direction taken by the agent. Their choices depending on the agent's position embody a policy map (top left).
doi:10.1371/journal.pcbi.1003024.g001

7. Critic implements Value



Fremaux et al. (2013)

7. Critic

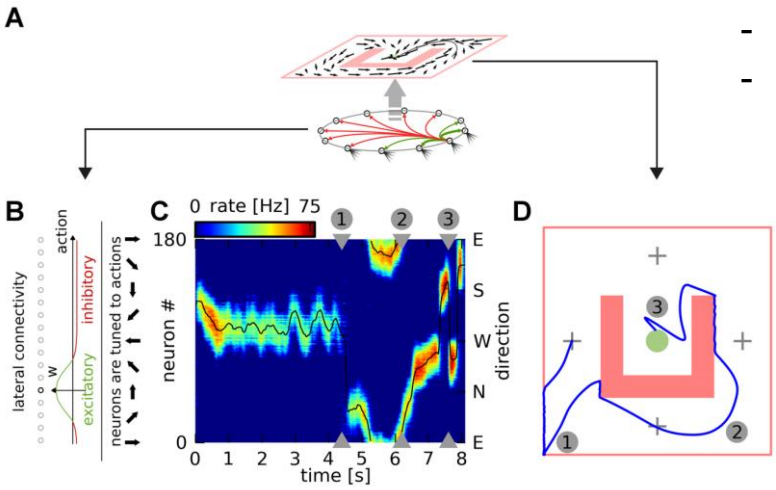
The task of the critic (previous slide)
B: Linear track task. The linear track experiment is a simplified version of the standard maze task. The actor's choice is forced to
the correct direction with constant velocity (left), while the critic learns to represent value (right).
C: Value function learning by the critic. Each colored
trace shows the value function represented by the critic neurons activity against time in the N~20 first simulation trials (from dark blue in trial 1 to dark red in trial 20), with t~tr corresponding to the time of the reward delivery. The black line shows an average over trials 30 to 50, after learning converged. The gray dashed line shows the theoretical value function.
D: TD signal d(t) corresponding to the simulation in C. The gray dashed line shows the reward time course r(t).

Fremaux et al. (2013)

7. Ring of Actor neurons implements policy

Note: no need to formally define a softmax function

- Local excitation
- Long-range inhibition
- Not a formal softmax



Fremaux et al. (2013)

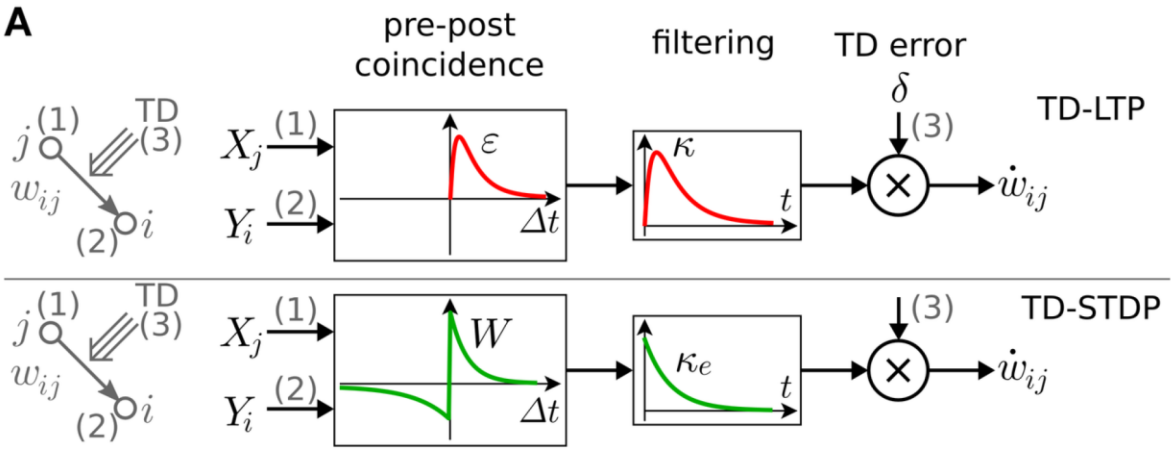
7. Ring of actor neurons

Actor neurons (previous slide).
A: A ring of actor neurons with lateral connectivity (bottom, green: excitatory, red: inhibitory) embodies the agent's policy (top).
B: Lateral connectivity. Each neuron codes for a distinct motion direction. Neurons form excitatory synapses to similarly tuned neurons and inhibitory synapses to other neurons.
C: Activity of actor neurons during an example trial. The activity of the neurons (vertical axis) is shown as a color map against time (horizontal axis). The lateral connectivity ensures that there is a single bump of activity at every moment in time. The black line shows the direction of motion (right axis; arrows in panel B) chosen as a result of the neural activity.
D: Maze trajectory corresponding to the trial shown in C. The numbered position markers match the times marked in C.

.

Fremaux et al. (2013)

7. Learning rule with TD in Actor-Critic for spiking neurons



Fremaux et al. (2013)

7. Learning rule with TD in Actor-Critic for spiking neurons

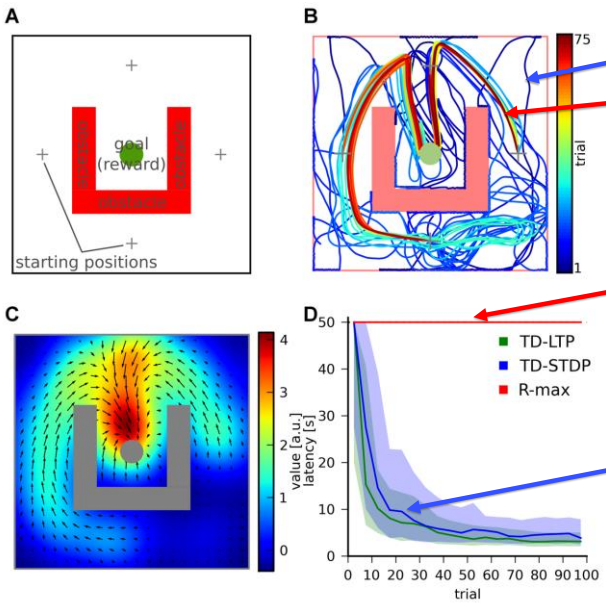
A: Learning rule with three factors (previous slide).

Top: TD-LTP is the learning rule resulting from policy gradient. It works by passing the presynaptic spike train X_j (factor 1) and the postsynaptic spike train Y_i (factor 2) through a coincidence window e. Spikes are counted as coincident if the postsynaptic spike occurs within after a few ms of a presynaptic spike. The result of the pre-post coincidence measure is filtered through a kernel (which yields the eligibility trace), and then multiplied by the TD error $\delta(t)$ (factor 3) to yield the learning rule which controls the change of the synaptic weight w_ ij .

Bottom: TD-STDP is a TD-modulated variant of R-STDP. The main difference with TD-LTP is the presence of a post-before-pre component in the coincidence window.

Fremaux et al. (2013)

7. Maze Navigation with TD in Actor-Critic with spiking neurons



R-max:
Policy gradient without the critic. The goal was never found within 50s.

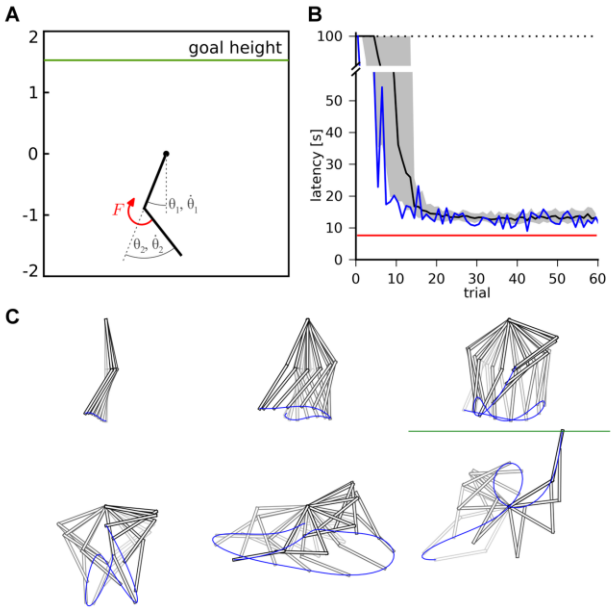
TD-STDP:
After 25 trials, the goal was found within 20s.

7. Maze Navigation with TD in Actor-Critic with spiking neurons

Maze navigation learning task.
A: The maze consists of a square enclosure, with a circular goal area (green) in the center. A U-shaped obstacle (red) makes the task harder by forcing turns on trajectories from three out of the four possible starting locations (crosses).
B: Color-coded trajectories of an example TD-LTP agent during the first 75 simulated trials. Early trials (blue) are spent exploring the maze and the obstacles, while later trials (green to red) exploit stereotypical behavior.
C: Value map (color map) and policy (vector field) represented by the synaptic weights of the agent of panel B after 2000s simulated seconds.
D: Goal reaching latency of agents using different learning rules. Latencies of N~100 simulated agents per learning rule. The solid lines shows the median shaded area represents the 25th to 75th percentiles. The R-max agent were simulated without a critic and enters times-out after 50 seconds.

Fremaux et al. (2013)

7. Acrobot task with TD in Actor-Critic with spiking neurons



Fremaux et al. (2013)

Previous slide.
Application of the same model to the Acrobot task.

7. TD in Actor-Critic with spiking neurons

- Learns in a few trials (assuming good representation)
- Works in continuous time.
- No artificial ‘events’ or ‘time steps’
- Works with spiking neurons
- Works in continuous space and for continuous actions
- Uses a biologically plausible 3-factor learning rule
- Critic implements value function
- TD signal calculated by critic
- Actor neurons interact via synaptic connections
- No need for algorithmic ‘softmax’

Fremaux et al. (2013)

Previous slide.
Summary of findings

Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited
5. Third factor
6. Actor-critic revisited
7. Example of Navigation
8. Generic example of action selection

Previous slide.
Can we go beyond one or two examples? What is the generic architecture?

review: Coarse Brain Anatomy and Reinforcement Learning

Actor-Critic Reinforcement learning needs:

- states / sensory representation → cortex?, hippocampus?

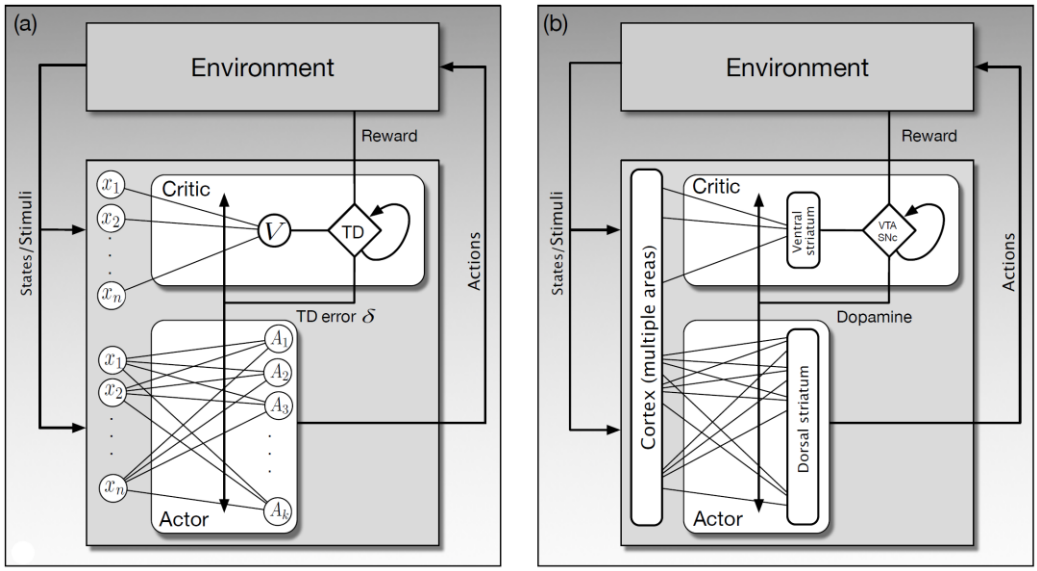
- action selection → striatum?, motor cortex?

- TD error signals → dopamine

- value function/critic → ?
- TD error calculation → ?

Previous slide.
Can we go beyond one or two examples? What is the generic architecture? What would we like to see in the brain? Wat do we need for RL?

8. Generic example for action selection by Actor-Critic



Book: Sutton and Barto 2018/Takahashi et al. 2008

8. Generic example for action selection by Actor-Critic

Actor-critic Artificial Neural Net (ANN) and a hypothetical neural implementation (previous slide)
a) Actor-critic algorithm as an ANN. The actor adjusts a policy based on the TD error it receives from the critic; the critic adjusts state-value parameters using the same . The critic produces a TD error from the reward signal, R, and the current change in its estimate of state values. The actor does not have direct access to the reward signal, and the critic does not have direct access to the action.
b) Hypothetical neural implementation of an actor-critic algorithm in the brain. The actor and the value-learning part of the critic are respectively placed in the ventral and dorsal subdivisions of the striatum. The TD error is transmitted by dopamine neurons located in the VTA and SubstantiaNigra (SNc) to modulate changes in synaptic efficacies of input from cortical areas to the ventral and dorsal striatum.
Adapted from Frontiers in Neuroscience, vol. 2(1), 2008, Y. Takahashi, G. Schoenbaum, and Y. Niv,

8. Summary

Several aspects of TD learning in an actor-critic framework can be mapped to the brain:

Sensory representation: Cortex and Hippocampu
Actor : Dorsal Striatum
Critic : Ventral Striatum (nucleus accumbens)

TD-signal: Dopamine

Learning in a few trials (not millions!) possible, if the sensory presentation is well adapted to the task

Previous slide. Summary

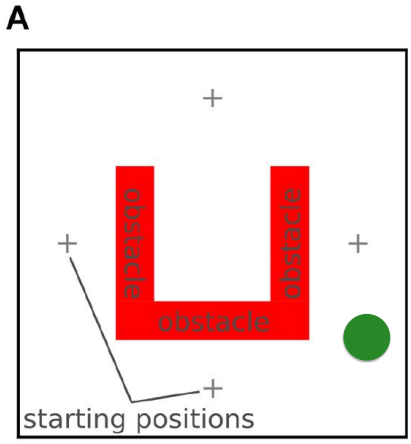
Artificial Neural Networks: Lecture 12
Reinforcement Learning and the Brain

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor learning rules
4. Policy gradient revisited
5. Third factor
6. Actor-critic revisited
7. Example of Navigation
8. Generic example of action selection
9. Model-based versus Model-free RL

Previous slide.
Final point: are we looking at the right type of RL algorithm?

9. Model-based versus Model-free

What happens in RL when you shift the goal after learning?



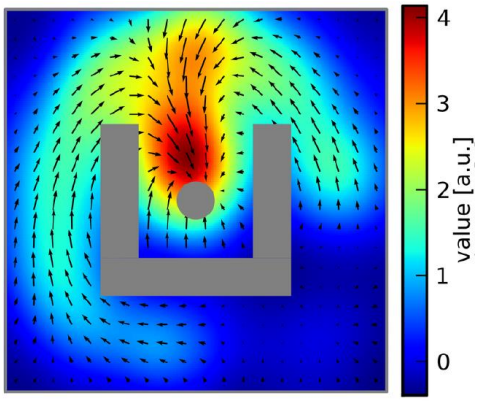
Previous slide.
Final point: are we looking at the right type of RL algorithm?
Imagine that the target location is shifted in the SAME environment.

9. Model-based versus Model-free Reinforcement Learning

What happens in RL when you shift the goal after learning?

→ The value function has to be re-learned from scratch.

agent learns ‘arrows’, but not the lay-out of the environment: Standard RL is ‘model-free’



Previous slide.
After a shift, the value function has to be relearned from scratch, because the RL algorithm does not build a model of the world. We just learn ‘arrows’: what is the next step (optimal next action), given the current state?

9. Model-based versus Model-free Reinforcement Learning

Definition:
Reinforcement learning is model-free, if the agent does not learn a model of the environment.

Note: of course, the learned actions are always implemented by some model, e.g., actor-critic.
Nevertheless, the term model-free is standard in the field.

Previous slide.
All standard RL algorithms that we have seen so far are ‘model free’.

9. Model-based versus Model-free Reinforcement Learning

Definition:
Reinforcement learning is model-based, if the agent does also learn a model of the environment.

Examples: Model of the environment
- state s1 is a neighbor of state s17.
- if I take action a5 in state s7, I will go to s8.
- The distance from s5 to s15 is 10m.
- etc

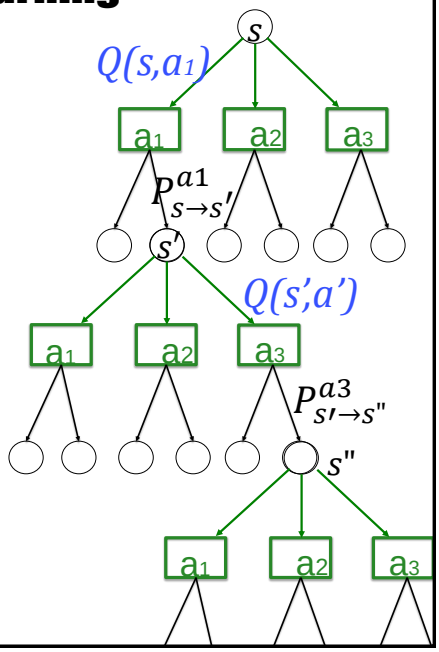
Previous slide.
Examples of knowledge of the environment, that would be typical for model based algorithm

9. Model-based versus Model-free Q-learning

Model-free:
the agent learns directly and only the Q-values

Model-based:
the agent learns the Q-values and also the transition probabilities

$P^{a1}_{s \to s'}$



Previous slide.
Let us go back to our 'tree'. If the algorithm knows the transition probabilities, then this means that it is a model-based algorithm

9. Model-based versus Model-free Reinforcement Learning

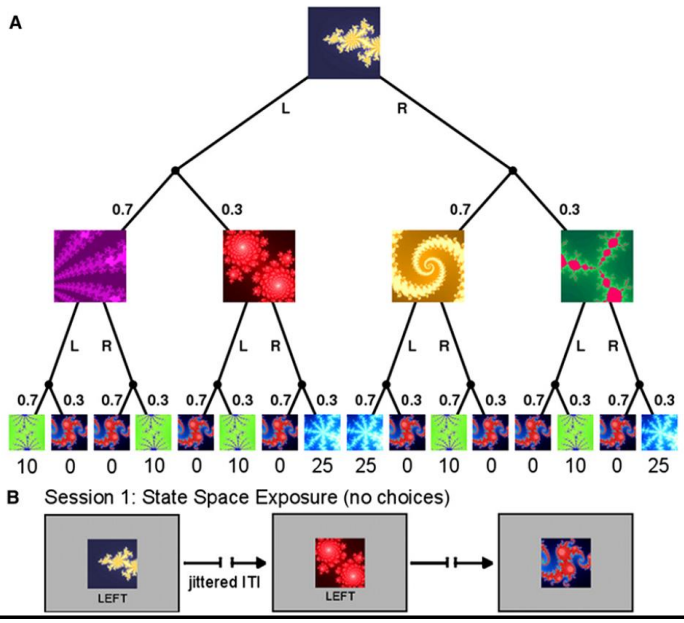
Advantages of Model-based RL:

- the agent can readapt if the reward-scheme changes
- the agent can explore potential future paths in its 'mind'
  → agent can plan an action path

Note: Implementations of Chess and Go are 'model-based', because the agent knows the rules of the game and can therefore plan an action path. It does not even have to learn the 'model'.

next slide.
The next question then is: do humans work with model-free or model-based RL?

9. Model-based learning



Model based:
You know what state to expect given current state and action choice.
'state prediction'

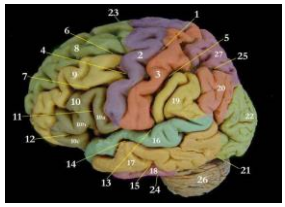
Gläscher et al. 2010

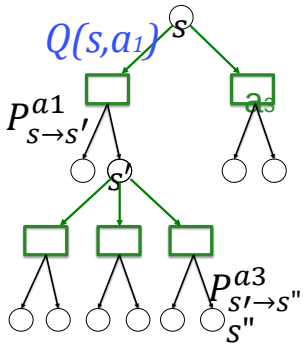
State and Reward Prediction Task (previous slide)
(A) The experimental task was a sequential two-choice Markov decision task in which all decision states are represented by fractal images. The task design follows that of a binary decision tree. Each trial begins in the same state. Subjects can choose between a left (L) or right (R) button press. With a certain probability (0.7/0.3) they reach one of two subsequent states in which they can choose again between a left or right action. Finally, they reach one of three outcome states associated with different monetary rewards (0, 10cent, and 25cent).

Gläscher et al. 2010

9. Model-based Reinforcement learning

$$Q(s,a) = \sum_{s'} P^{a}_{s \to s'} \left[R^{a}_{s \to s'} + \gamma \sum_{a'} \pi(s',a') Q(s',a') \right]$$





Gläscher et al. 2010

(previous slide)

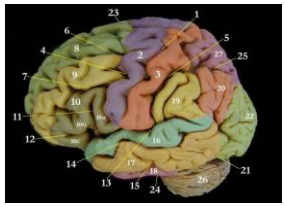
) Between scanning sessions subjects were presented with the reward schedule that maps the outcome states to the monetary payoffs. This mapping was rehearsed in a short choice task.

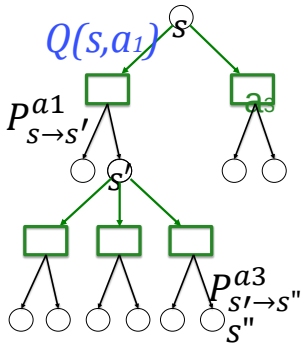
Gläscher et al. 2010

9. Model-based Reinforcement learning

Session 2: subject takes action

$$Q(s,a) = \sum_{s'} P^{a}_{s \to s'} \left[R^{a}_{s \to s'} + \gamma \sum_{a'} \pi(s',a') Q(s',a') \right]$$





State prediction error:
with p=0.3 you go to an 'unexpected' next state

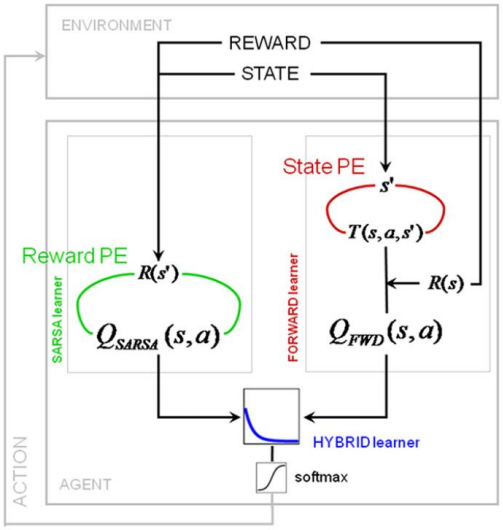
Gläscher et al. 2010

(previous slide)

Finally, in the second scanning session, subjects were free to choose left or right actions in each state. In addition, they also received the payoffs in the outcome states.

Gläscher et al. 2010

9. Model-free versus Model-based Reinforcement learning



After observed state transition, update

T(s,a,s') = $P^{a}_{s \to s'}$

Matrix of transition probabilities

'state predition error'

After observed reward, update

Q(s,a)

'reward predition error'

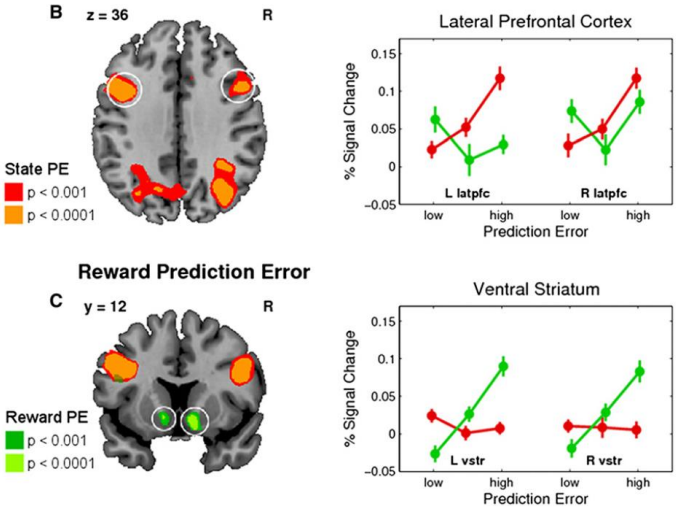
Gläscher et al. 2010

(previous slide)

Schematic of the information flow in model-free and model-based algorithms. In model-free algorithms we just update the Q-value. In model based algorithms, we also update an estimate of the transition probabilities.

Gläscher et al. 2010

9. State prediction error and reward prediction error in the brain



State prediction error significant in **lateral prefrontal cortex**

Reward prediction error significant in **ventral striatum** (includes nucleus accumbens)

Gläscher et al. 2010

(previous slide)
Dbrain areas that respond to model-based updates (state-prediction errors)
Are distinct from those that respond to the TD-like reward-prediction error

Gläscher et al. 2010

9. State prediction error and reward prediction error in the brain

- Certain brain areas seem to be involved in reward prediction
- Other brain areas seem to be in involved in state prediction
- Reinforcement Learning has influenced human brain science

Conclusion is NOT: Brain implements SARSA or Actor-critic or Q-learning.
Conclusion is modest: you can find correlations with signatures of RL in the brain.

(previous slide)
Summary of experiment

Gläscher et al. 2010

Summary

Learning outcome for today:

- three-factor learning rules can be implemented by the brain

→ synaptic changes need presynaptic factor, postsynaptic factor and a neuromodulator (3rd factor)

→ actor-critic and other policy gradient methods give rise to very similar three-factor rules

- **eligibility traces as 'candidate parameter updates'**

→ set by joint activation of pre- and postsynaptic factor

→ decays over time

→ transformed in weight update if dopamine signal comes

- the dopamine signal has signature of the TD error

→ responds to reward minus expected reward

→ responds to unexpected events that predict reward

- **Signatures of model-based and model-free RL in the brain**

Thanks!

And good luck for the exam