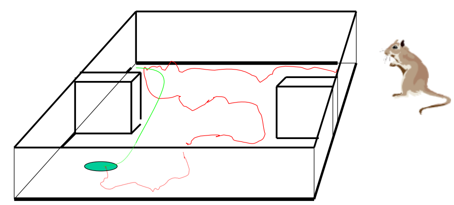
Neural Networks and Biological Modeling: Lecture 6

Introduction to Reinforcement Learning

- **-Learning by reward**
- **-Reward based learning (at the synapse)**
- **-Detour: rat navigation**
  - -Place cells and Rat hippocampus
  - -A model of spatial representation
  - -Learning to find the goal location
- **-Reinforcement learning theory**
  - - 1-step horizon
  - - multi-step action sequences
- **-Eligibitiy traces**
- **-Model of Rat navigation**
  - - behavioral experiments

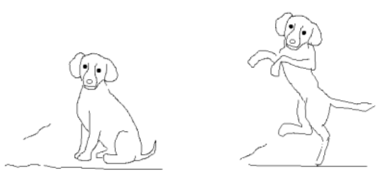
Wulfram Gerstner, EPFL

Learning by reward



Rat learns to find the reward

Learning by reward: conditioning

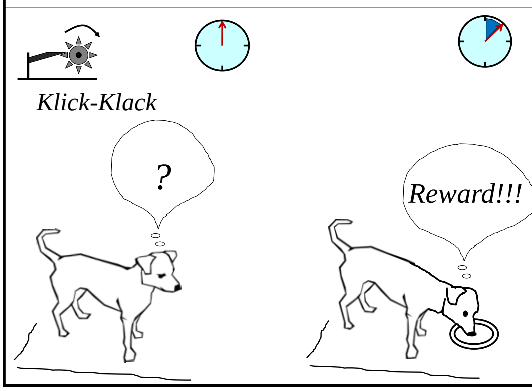


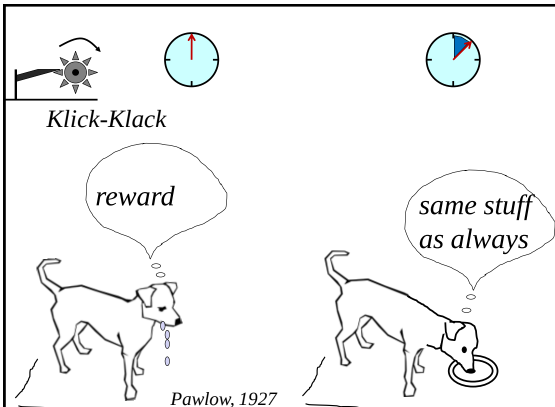
*The dog likes to do things, for which it gets **reward**!*

Trial and error learning

Instrumental conditioning: action learning

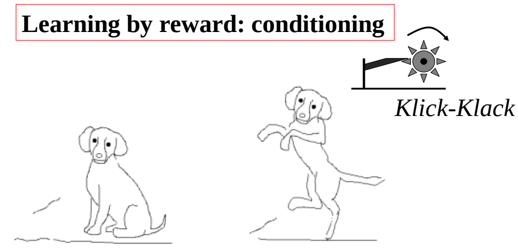
Detour: Classical Conditioning *Pawlow, 1927*





Pawlow, 1927

Learning by reward: conditioning



*The dog likes to do things, for which it gets a **Klick-Klack!***

Trial and error learning

(the Klick-Klack replaces the explicit reward)

Learning by reward: humans?

What is the reward?

Learning actions, skills, tricks, by reward

Images taken from public WEB pages

Neural Networks and Biological Modeling: Lecture 6

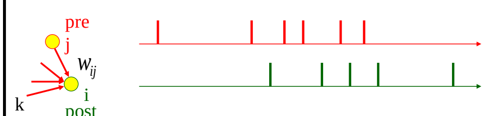
Introduction to Reinforcement Learning

- **-Learning by reward**
- **-Reward based learning (at the synapse)**
  - -Hebbian learning/unsupervised
  - -reinforcement learning
- **-Detour: rat navigation**
  - -Place cells and Rat hippocampus
  - -A model of spatial representation
  - -Learning to find the goal location
- **-Reinforcement learning theory**
- **-Eligibitiy traces**
- **-Model of Rat navigation**
  - behavioral experiments

Wulfram Gerstner, EPFL

Unsupervised vs. reinforcement learning

review: Hebbian Learning

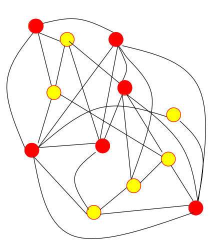


*When an axon of cell **j** repeatedly or persistently takes part in firing cell **i**, then j's efficiency as one of the cells firing i is increased*

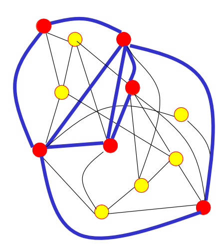
Hebb, 1949

- local rule
- simultaneously active (correlations)

Hebbian Learning



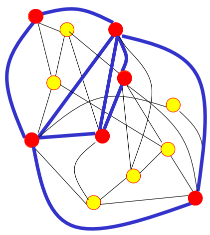
Hebbian Learning



item memorized

Hebbian Learning

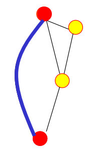
Recall:
Partial info



item recalled

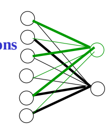
Associative memory
Lecture 5

Hebbian Learning
= unsupervised learning



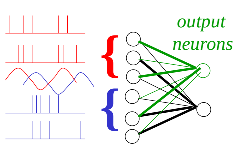
Detour: Receptive field development

Initial:
random
connections

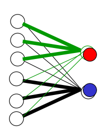


unselective
neurons

Hebbian leanring leads to specialized
Neurons (developmental learning);
(Lecture 4)

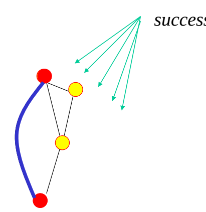


Correlated input



output neurons specialize

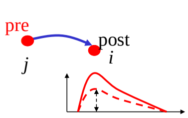
Reinforcement Learning
= reward + Hebb



Classification of plasticity: unsupervised vs reinforcement

LTP/LTD/Hebb

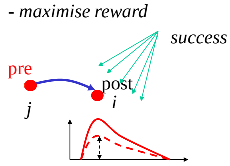
Theoretical concept
- *passive changes*
- *exploit statistical correlations*



Functionality
-useful for development
(wiring for receptive fields)

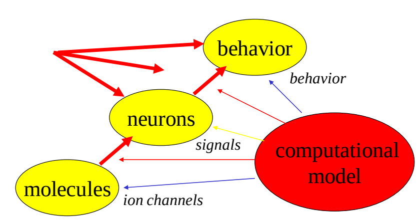
Reinforcement Learning

Theoretical concept
- *conditioned changes*
- *maximise reward*



Functionality
- *useful for learning*
a new behavior

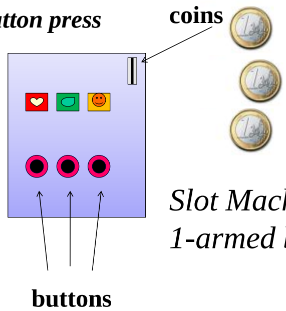
Learning by reward / reinforcement learning



Swiss Federal Institute of Technology Lausanne, EPFL
Laboratory of Computational Neuroscience, LCN, CH 1015 Lausanne

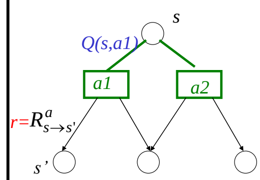
Learning by reward / reinforcement learning

action=button press



Slot Machine
1-armed bandid

Reward-based Action Learning



s=state

a2 *a=action*

$R^a_{s\to s'}$ = *reward*

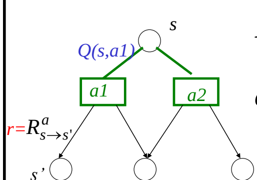
s'=new action

Q(s,a)=expected reward

$Q(s,a)=\sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Blackboard:

Reward-based Action Learning



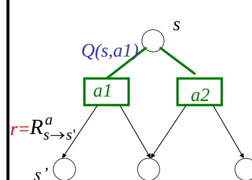
Best action:

a* = max$_a$ Q(s,a)

Q(s,a)=expected reward

$Q(s,a)=\sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Action choice: Policy π



Best action:

a = max$_a$ Q(s,a)*

(if Q values are known)

Good policy π (ε –greedy)

a = max$_a$ Q(s,a) with prob 1-ε*

arbitrary with prob ε

(if Q values are estimated)

Q(s,a)=expected reward

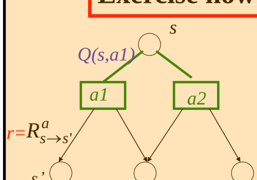
$Q(s,a)=\sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Estimate Q iteratively

Iterative Update

$\Delta Q(s,a)=\eta\ [r-Q(s,a)]$

Exercise now: Iterative update



s=state

a2 *a=action*

$R^a_{s\to s'}$ = *reward*

s'=new action

a)Show that an empirical evaluation of Q(s,a) by averaging the the rewards for action a over k or (k+1) trials, leads to an iterative update rule of the form $\Delta Q(s,a)=\eta\ [r-Q(s,a)]$

b) Calculate eta.

c) Give an intuitive explanation of the update rule

expected reward

$Q(s,a)=\sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Next lecture** at **10h15

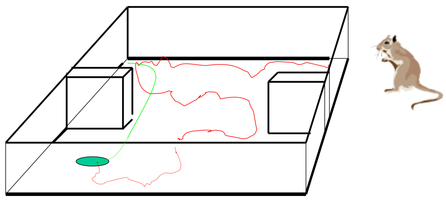
Neural Networks and Biological Modeling: Lecture 6

Introduction to Reinforcement Learning

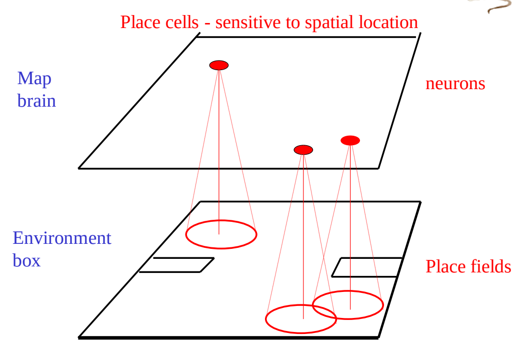
- **-Learning by reward**
- **-Reward based learning**
- **-Detour: rat navigation**
  - -Place cells and Rat hippocampus
  - -A model of spatial representation
  - -Learning to find the goal location
- **-Reinforcement learning theory**
- **-Eligibitiy traces**
- **-Model of Rat navigation**
  - behavioral experiments

Wulfram Gerstner, EPFL

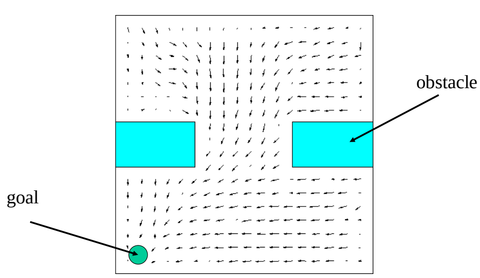
Biological Principles of Learning: spatial learning



Spatial representation



Use map



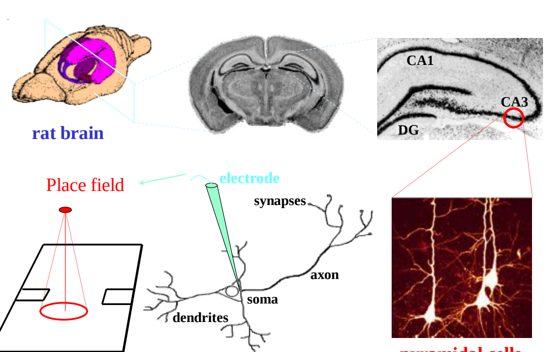
SPATIAL REPRESENTATION

- ***Model of place cells***

GOAL LEARNING

- ***Reward-based learning system***

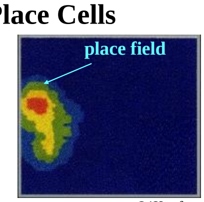
Neurophysiology of the Rat Hippocampus



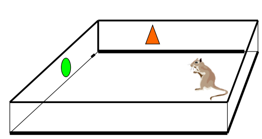
Hippocampal Place Cells

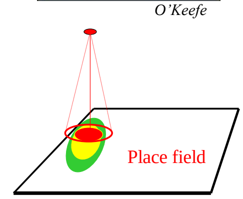
Depends on

- visual cues
- works also in the dark

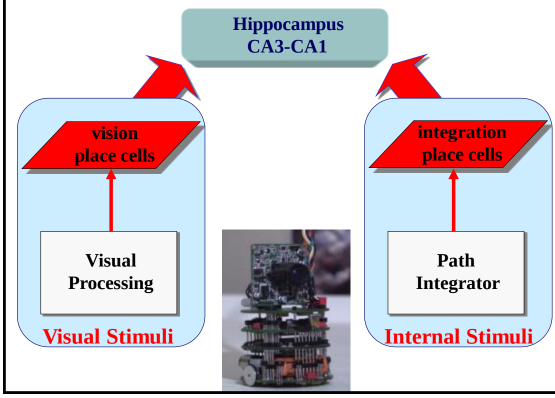


O'Keefe

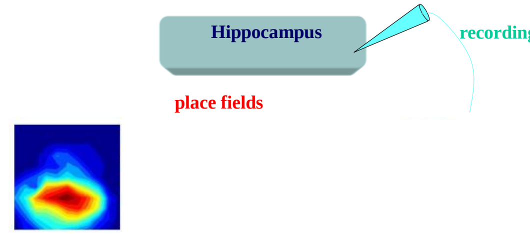




The Neural Model: Place Cells



The Neural Model: Place Cells



√ *SPATIAL REPRESENTATION*

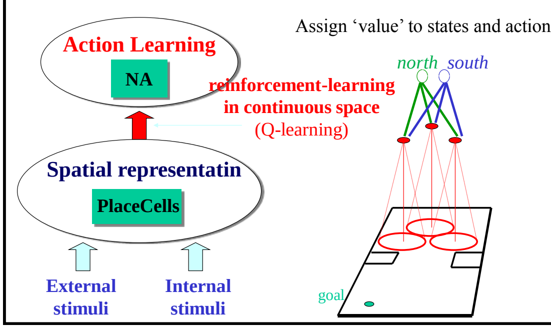
- ***Model of place cells***

→ *GOAL LEARNING*

- ***Reward-based learning system***

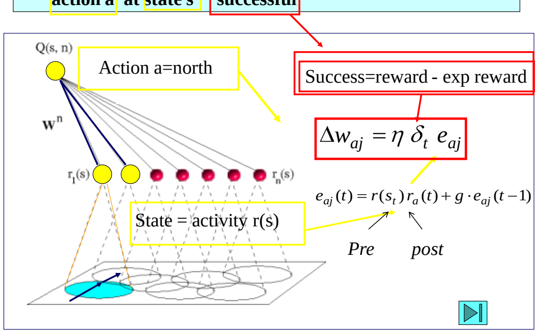
Reward-based Action Learning

Hippocampal place cells = fuzzy discretisation of continuous space



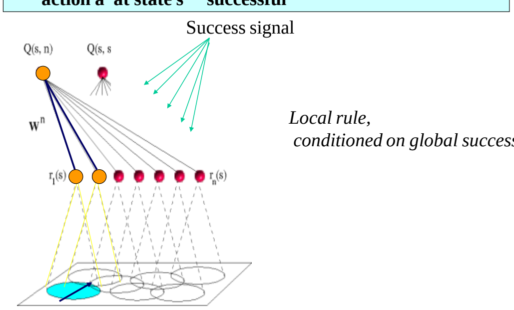
Reward-based Action Learning

Connection reinforced if
action a at state s successful



Reward-based Action Learning

Connection reinforced if
action a at state s successful



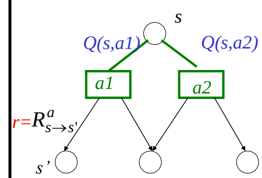
Neural Networks and Biological Modeling: Lecture 6

Introduction to Reinforcement Learning

-Learning by reward
-Detour: rat navigation
-Place cells and Rat hippocampus
-A model of spatial representation
-Learning to find the goal location
-Reward based learning
-Reinforcement learning theory
- one-step horizon
- multiple-step action sequences
-Eligibitiy traces
-Model of Rat navigation
behavioral experiments

Wulfram Gerstner, EPFL

Reward-based Action Learning



If you know Q(s,a), best action choice:

$a_t^* = \arg\max_a Q_a(s,a)$

Iterative Update

$\Delta Q(s,a) = \eta\ [r - Q(s,a)]$

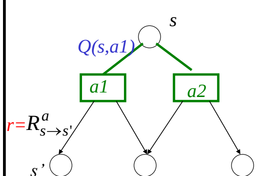
Exploration/exploitation dilemma

Q(s,a)=expected reward

$Q(s,a) = \sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Reward-based Action Learning

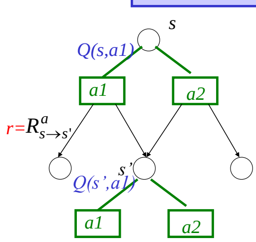
Blackboard:



$Q(s,a) = \sum_{s'} P^a_{s\to s'} R^a_{s\to s'}$

Update

$\Delta Q(s,a) = \eta\ [r - Q(s,a)]$



$\Delta Q(s,a) = \eta\ [r - (Q(s,a) - Q(s',a'))]$

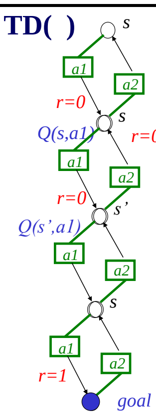
Reward-based Learning TD( )

- ***SARSA***

$\Delta Q(s,a) = \eta\ [r - (Q(s,a) - Q(s',a'))]$

- ***policy for action choice:***
Pick most often action

$a_t^* = \arg\max_a Q_a(s,a)$



Exercise now

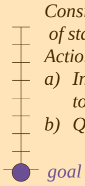
- ***Update of Q values in SARSA***

$\Delta Q(s,a) = \eta\ [r - (Q(s,a) - Q(s',a'))]$

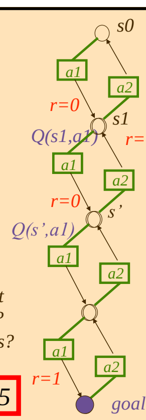
- ***policy for action choice:***
Pick most often action

$a_t^* = \arg\max_a Q_a(s,a)$

Consider a linear sequence of states. Reward only at goal. Actions are up or down.
a) Initialise Q values at 0. Start at top. How do Q values develop?
b) Q values after 3 complete trials?



Next lecture at 11:15

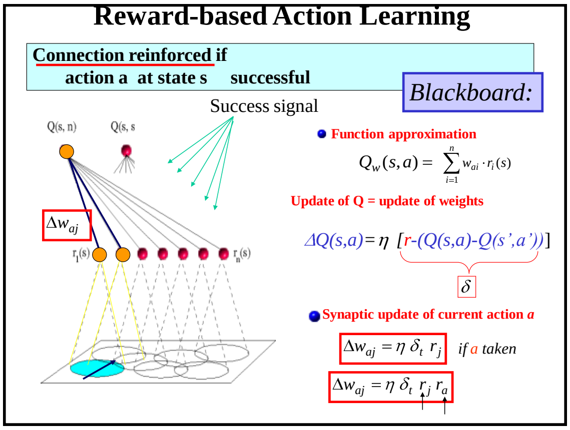


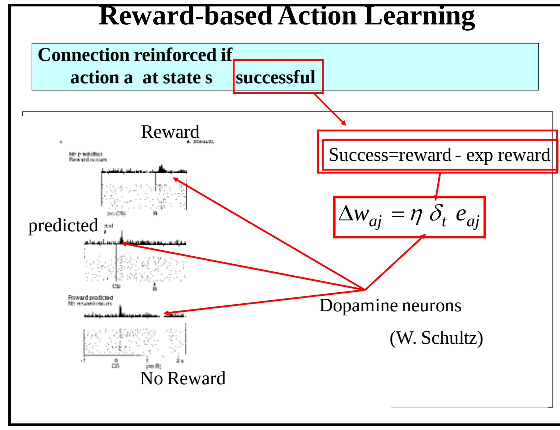
Neural Networks and Biological Modeling: Lecture 6

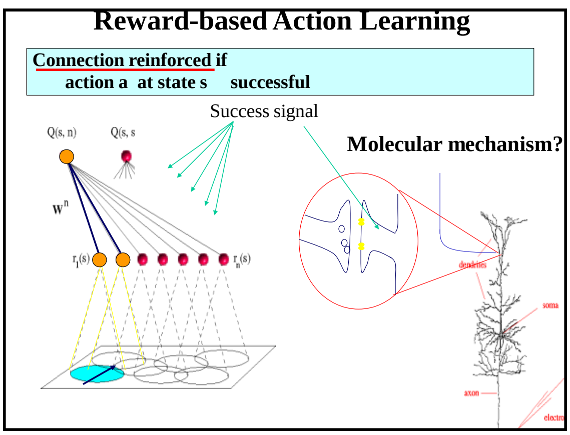
Introduction to Reinforcement Learning

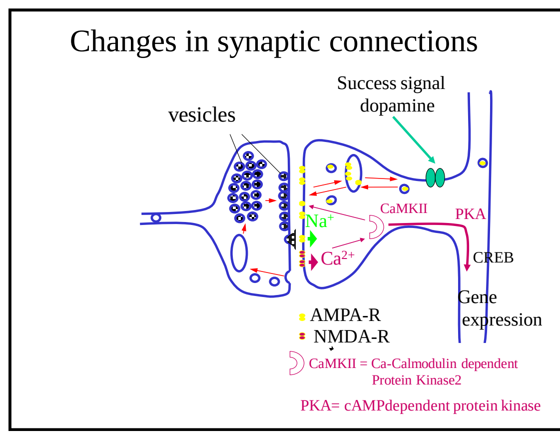
-Learning by reward
-Detour: rat navigation
-Place cells and Rat hippocampus
-A model of spatial representation
-Learning to find the goal location
-Reinforcement learning theory
- one-step horizon
- multiple-step action sequences
- continuous state space
-Eligibitiy traces
-Model of Rat navigation
behavioral experiments

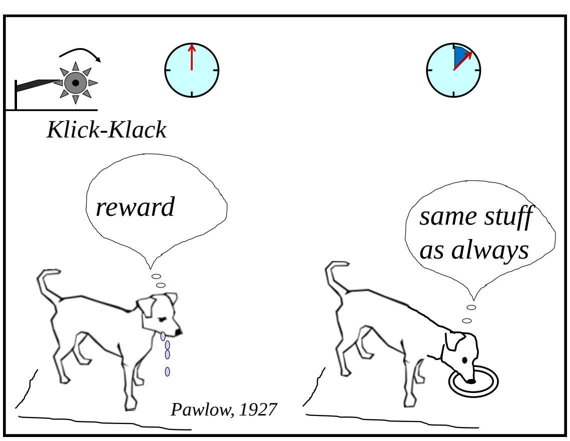
Wulfram Gerstner, EPFL

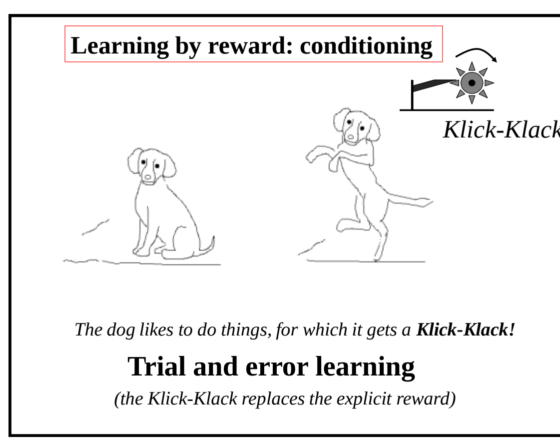

Reward-based Action Learning
Connection reinforced if
action a at state s successful
Blackboard:
Success signal
Q(s, n)
Q(s, s
Δw_{aj}
$r_1(s)$
$r_n(s)$
Function approximation
$Q_w(s,a) = \sum_{i=1}^{n} w_{ai} \cdot r_i(s)$
Update of Q = update of weights
$\Delta Q(s,a) = \eta\ [r-(Q(s,a)-Q(s',a'))]$
δ
Synaptic update of current action a
$\Delta w_{aj} = \eta\ \delta_t\ r_j$ if a taken
$\Delta w_{aj} = \eta\ \delta_t\ r_j\ r_a$


Reward-based Action Learning
Connection reinforced if
action a at state s successful
Reward
predicted
No Reward
Success=reward - exp reward
$\Delta w_{aj} = \eta\ \delta_t\ e_{aj}$
Dopamine neurons
(W. Schultz)


Reward-based Action Learning
Connection reinforced if
action a at state s successful
Success signal
Q(s, n)
Q(s, s
w^n
$r_1(s)$
$r_n(s)$
Molecular mechanism?


Changes in synaptic connections
vesicles
Success signal
dopamine
Na^+
Ca^{2+}
CaMKII
PKA
CREB
Gene expression
AMPA-R
NMDA-R
CaMKII = Ca-Calmodulin dependent Protein Kinase2
PKA= cAMPdependent protein kinase


Klick-Klack
reward
same stuff as always
Pawlow, 1927


Learning by reward: conditioning
Klick-Klack
The dog likes to do things, for which it gets a Klick-Klack!
Trial and error learning
(the Klick-Klack replaces the explicit reward)

The end

.... Of the lecture on Monday 28.March 2011,

.... Remaining slides next Monday

Neural Networks and Biological Modeling: Lecture 6

Introduction to Reinforcement Learning

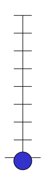
-Learning by reward
-Detour: rat navigation
-Place cells and Rat hippocampus
-A model of spatial representation
-Learning to find the goal location
-Reinforcement learning theory
- one-step horizon
- multiple-step action sequences
- continuous state space
-Eligibitiy traces
-Model of Rat navigation
behavioral experiments

Wulfram Gerstner, EPFL

Problem: learning is slow

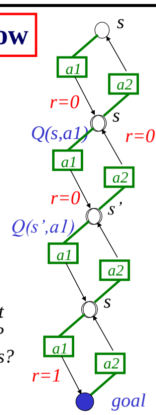
- Slow diffusion of information across several states

$$a_t^* = \arg\max_a Q_a(s,a)$$

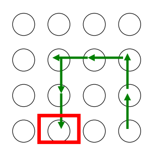


Consider a linear sequence of states. Reward only at goal. Actions are up or down.
a) Initialise Q values at 0. Start at top. How do Q values develop?
b) Q values after 3 complete trials?

goal

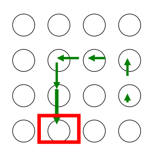


Eligibitiy trace



s=state
a=action
$R^a_{s\to s'}$ *= reward*
s'=new state

Sarsa $\Delta Q(s,a) = \eta\ [r - (Q(s,a) - \gamma\, Q(s',a'))]$



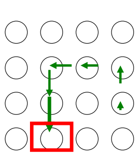
Idea: at the moment of reward update also previous action values along trajectories

Eligibitiy trace - algo

Sarsa(0) $\Delta Q(s,a) = \eta\ [r - (Q(s,a) - \gamma\, Q(s',a'))]$

$$\delta_t = r_{t+1} - Q_a(s,a) - \gamma \cdot Q_a(s',a')$$

Sarsa(λ) $\boxed{\Delta Q_{aj} = \eta\ \delta_t\ e_{aj}}$



if $a =$ *action taken in state j*
$e_{aj}(t) = \gamma\lambda\, e_{aj}(t-\Delta t) + 1$
else
$e_{aj}(t) = \gamma\lambda\, e_{aj}(t-\Delta t)$

Reward-based Learning TD( )

- ***TD error in SARSA***

$$\delta_t = R_{t+1} - Q_a(s,a) - \gamma \cdot Q_a(s',a')$$

- **Function approximation**

$$Q_w(s,a) = \sum_{i=1}^{n} w_i^a \cdot r_i(s)$$

- ***policy for action choice:***
Pick most often action

$$a_t^* = \arg\max_a Q_a(s,a)$$

- **Synaptic update**

$$\boxed{\Delta w_{aj} = \eta\ \delta_t\ e_{aj}}$$

pre post

- ***Eligibility trace (memory at synapse):***

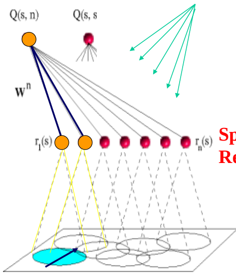
pre post

$$e_{aj}(t) = \gamma\lambda\, e_{aj}(t-\Delta t) + \begin{cases} r_j\, r_a & \text{if } a = \text{action taken} \\ 0 & \end{cases}$$

memory

Reward-based Action Learning

Connection reinforced if
action a at state s successful



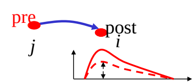
Learning Rule

$$\Delta w_{aj} = \eta \, \delta_t \, e_{aj}$$

Spatial Representation

Plasticity models: unsupervised vs reinforcement

STDP/Hebb



Model

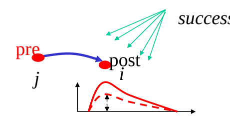
- *STDP (see above)*

$$\Delta w_{ij} \propto pre \cdot post + other\ terms$$

Reinforcement Learning

theoretical Protocol

- *maximise reward*



Model

- *spike-based model?*

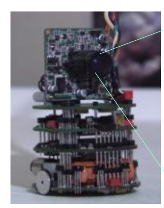
$$\Delta w_{ij} \propto success \cdot (pre \cdot post + other\ terms)$$

Introduction to reinforcement learning (via a model of rat navigation)

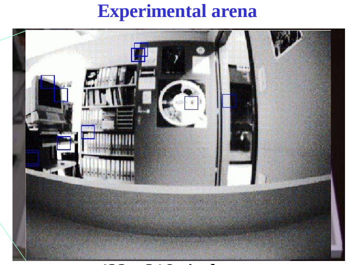
-Basics or rat navigation
-Place cells and Rat hippocampus
-A model of spatial representation
-Learning to find the goal location
-Reward based learning (basic ideas)
-Reinforcement learning theory
-Eligibitiy traces
-the full model: behavioral experiments

Validating the Model

The KHEPERA mobile miniature robot



Experimental arena

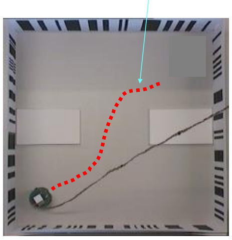


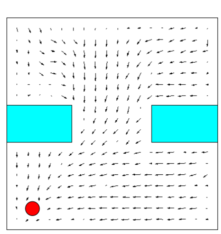
422 x 316 pixels

Open-field Navigation Experiments

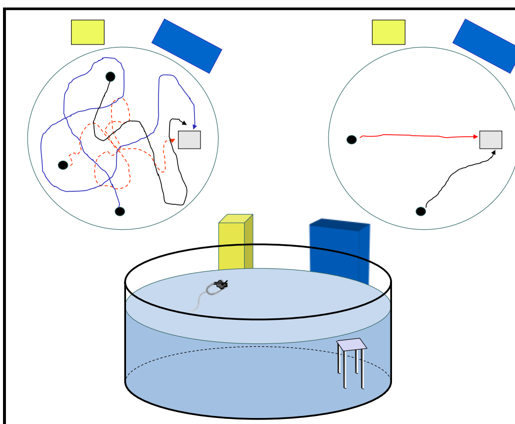
robot trajectory

(Biol. Cybern., 2000)



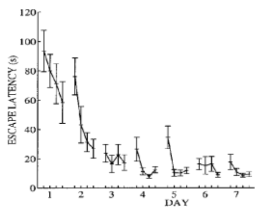


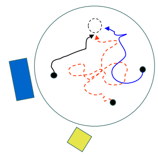
Navigation map
after 20 training trials



Behavior: Navigation to a hidden goal (Morris water maze)

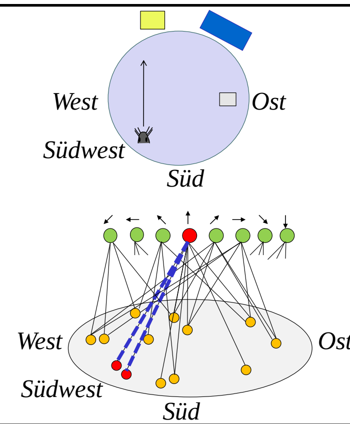
- *Different starting positions*
- *Task learning depends on the hippocampus*

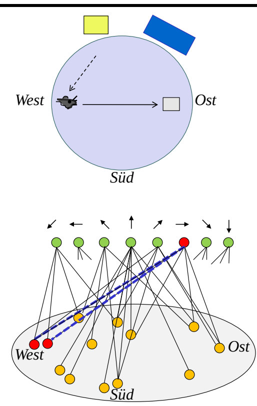


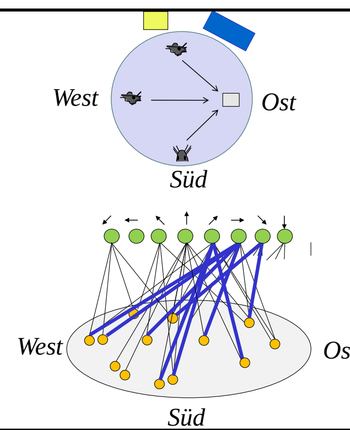


61

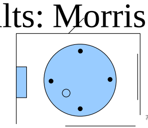
Foster, Morris & Dayan 2000

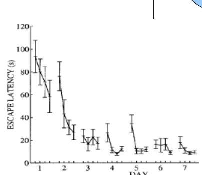


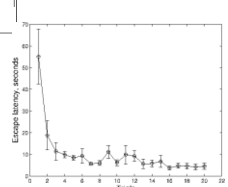




Results: Morris water maze







Experiments

Foster, Morris & Dayan 2000

Model,
Sheynikhovich et al.,
July 2009

65

Exercise 3.1 now; Computer demo at 12:00

- ***Update of Q values in SARSA***

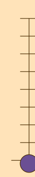
  $\Delta Q(s,a) = \eta\ [r-(Q(s,a)-Q(s',a'))]$

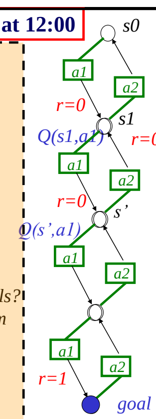
- ***policy for action choice:***

  Pick most often action

  $a_t^* = \arg\max_a Q_a(s,a)$

  Consider a linear sequence of states. Reward only at goal.
  Use eligibility trace
  a) Q values after 2 complete trials?
  b) Increase number of states from n=4 to 8; rescaling?





The end

I hope you enjoyed the show ....

.... See you Monday