Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 1: Examples of Reward-based Learning

Objectives for Lecture RL1 (Part 1-3)

- Reinforcement Learning (RL) is learning by rewards
- Agents and actions, states and rewards
- Convergence in expectation, online and batch.

Reading:
Sutton and Barto, Reinforcement Learning (MIT Press, 2nd edition 2018)
Chapters: 1.1-1.4; 2.1-2.6; 3.1-3.5; 6.4

Reading for this week:

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

Chapters: 1.1-1.4; 2.1-2.6; 3.1-3.5; 6.4

Background reading:

Silver et al. 2017, Archive

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

Artificial Neural Networks for action learning

No labeled data?

Replaced by:

‘Value of action’

- ‘goodie’ for dog
- ‘success’
- ‘compliment’

BUT:

Reward is rare:

‘sparse feedback’ after a long action sequence

Previous slide.

How does a human learn to play table tennis: How does a child learn to play the piano? How does a dog learn to perform tricks?

In all these cases there is no supervisor. No master guides the hand of the players during the learning phase. Rather the player ‘discovers’ good movements by rather coarse feedback. For example, the ball in table tennis does not land on the table as it should. That is bad (negative feedback). The ball has a great spin so that the opponent does not get. This is good (positive feedback).

Similarly, it is hard to tell a dog what to do. But if you reinforce the dog’s behavior by giving a ‘goodie’ at the moment when it spontaneously performs a nice action, then it can learn quite amazing things.

In all these cases it is the ‘reward’ that guides the learning. Rewards can be the goodie for the dog, or just the feeling ‘now I did well’ for humans.

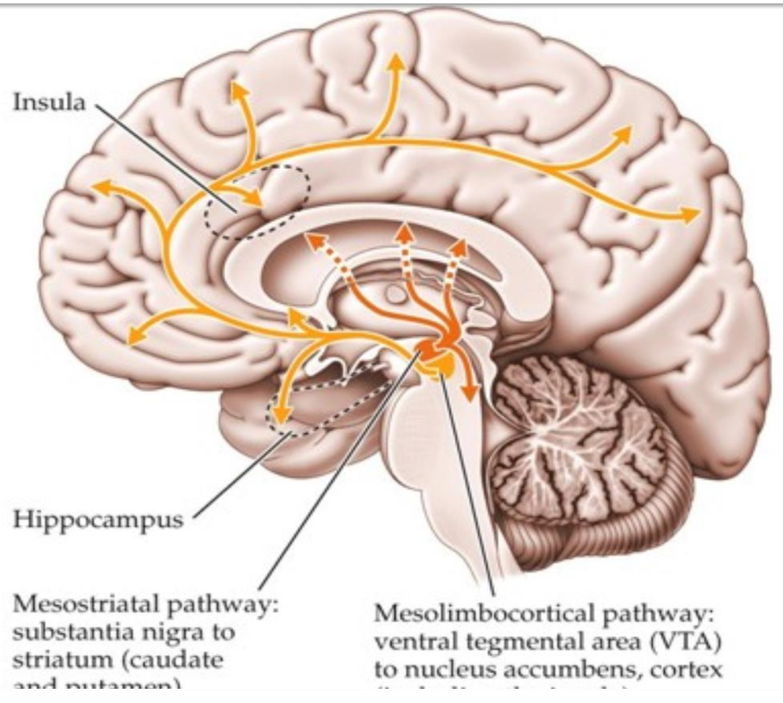
Reward information is available in the brain

Neuromodulator **dopamine**:
Signals “reward minus expected reward”

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

‘success signal’

Dopamine



Previous slide.

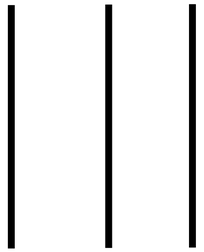
Inside the brain, reward information is transmitted by the neuromodulator dopamine. Neurons that use dopamine as their chemical transmission signal are situated in nuclei below the cortex and have cables (axons) that reach out to vast areas of the brain.

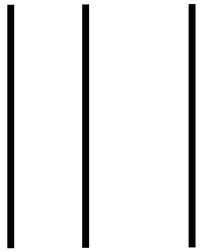
As we will see later, neurons that communicate with the neuromodulator dopamine transmit a generic success signal that is not just reward, but something like ‘reward minus expected reward’.

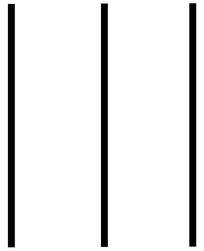
To conclude, reward information is available throughout the brain.

Examples of reinforcment learning

Middle bar: shifted left or shifted right?

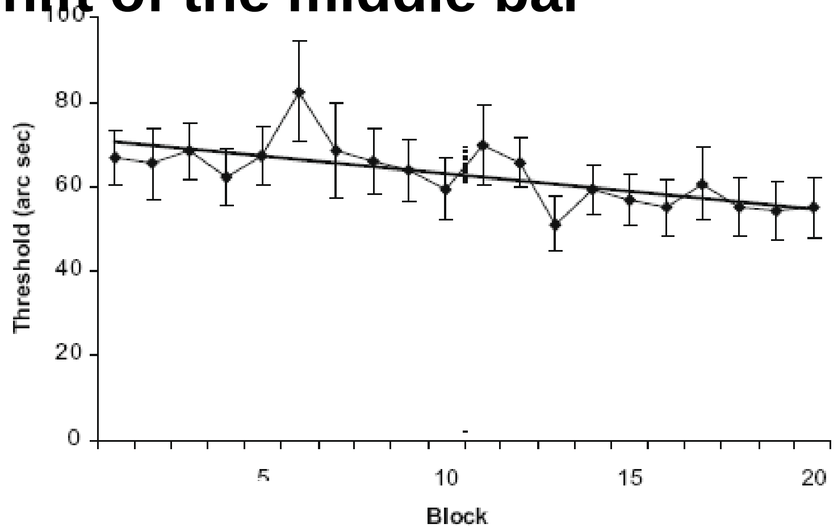






Observers get better at seeing the shift of the middle bar

Min. shift



Feedback:
tone for wrong response

Tartaglia,Aberg,Herzog 2009

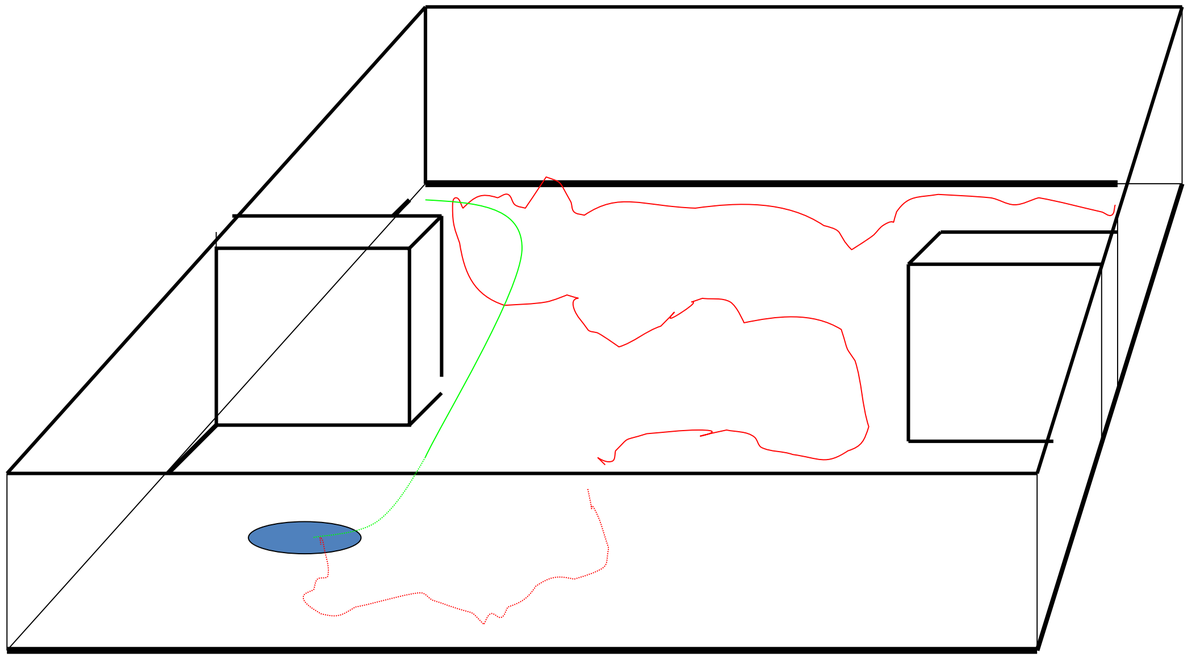
Previous slide (This example is not shown in class)

Let us look at a few additional examples, beyond table tennis.

Humans can get, by practice and feedback, better at recognizing a visual pattern with three bars. The task is to distinguish cases where the middle bar is shifted to the left from those where it is shifted to the right.

Bottom right:
The minimal shift that is just recognizable decreases over time (1 block = 1 practice session) indicating learning.
The feedback signal is just right or wrong.

Examples of reinforcement learning: animal conditioning

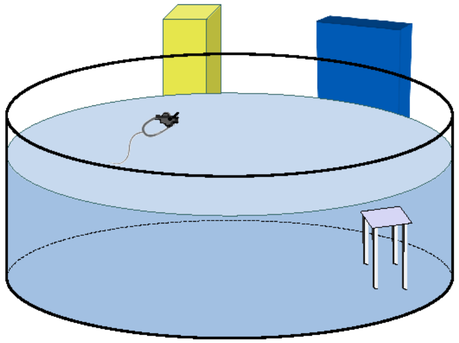


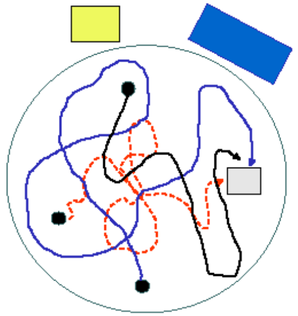
Previous slide.

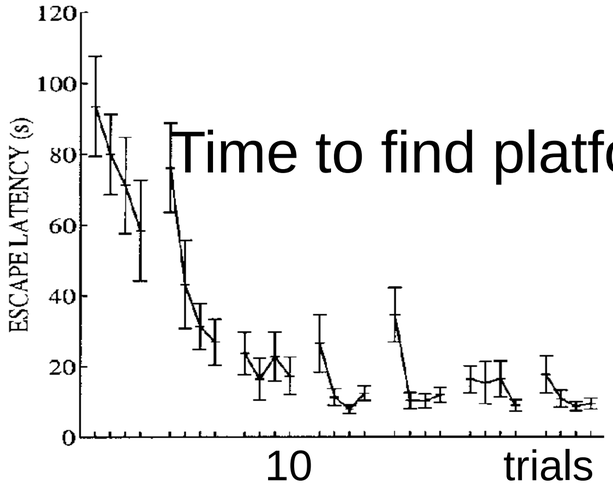
If you put a rat into an environment it will wander around. Suppose that, at some place, it discovers a food source hidden below the sand of the surface.
After a couple of trials it will go straight to the location of the food source which implies that it has learned the appropriate sequence of actions in the environment to find the food source.

Examples of reinforcement learning: 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

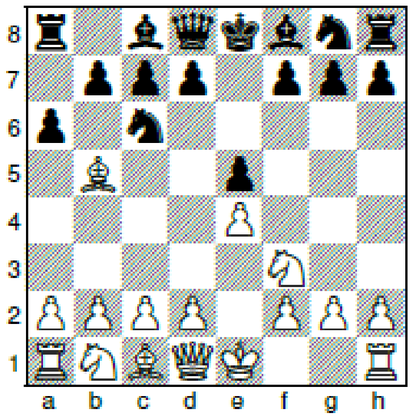
Previous slide.

Actual experiments for location learning are often performed in a Morris water maze. In the maze, there are 4 starting points and one target location which is a platform hidden (in milky water) just below the water surface. The rat does not like to swim in cold water and therefore tries to find the platform.

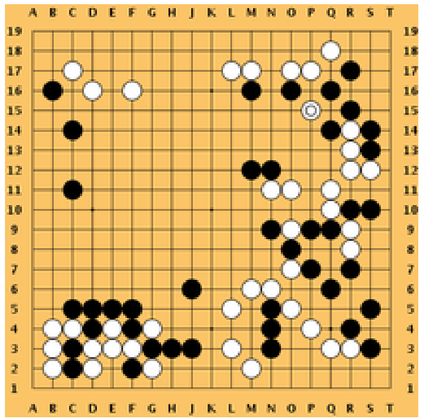
After a few trials it swims straight to the platform.
Bottom right: the time to reach the platform decreases over trials, indicating learning.

Deep reinforcement learning

Chess



Go



Artificial neural network (*AlphaZero*) discovers different strategies by playing against itself.

In Go, it beats Lee Sedol

Previous slide.

In chess a neural network trained by reinforcement learning discovers winning strategies by playing against itself. Similarly, a neural network playing Go against itself learns to play at a level so as to beat one of the world champions.

The aim of the class is to arrive at Deep Reinforcement Learning (Deep RL): Today we start with (standard) RL, in a few weeks we turn to deep networks, and in May we will turn to Deep RL.

Deep reinforcement learning

Network for choosing action

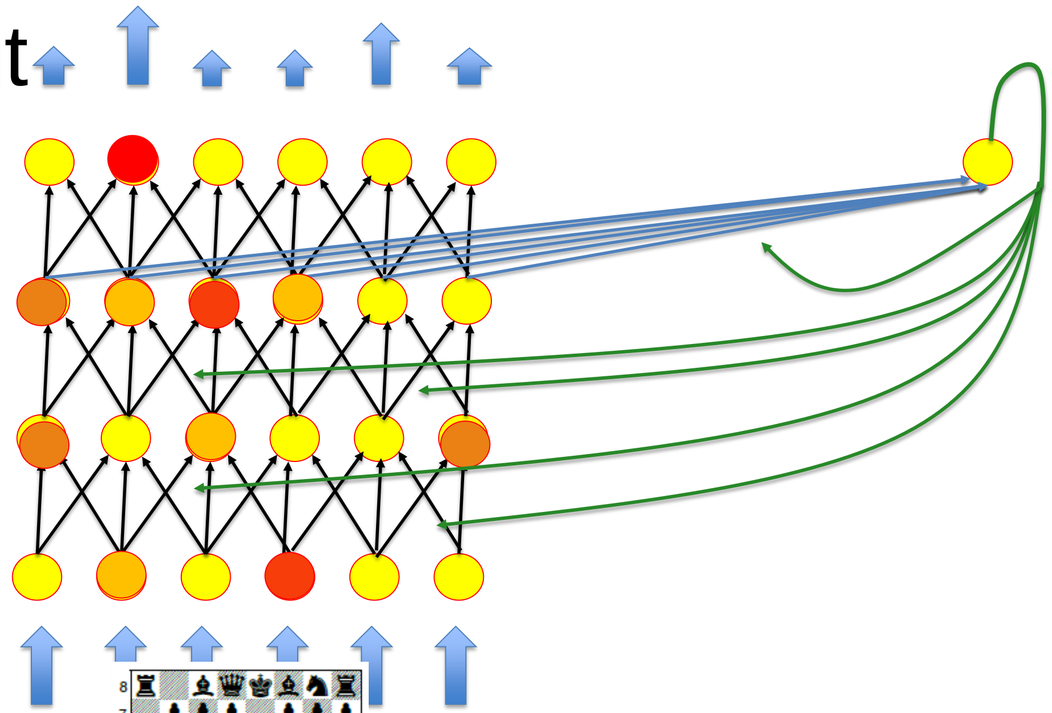
action:

Advance king

2nd output for **value** of state:

probability to win

output



input

Learning by success signal

- change connections

aim:

- choose next action to win

aim for value unit:

- predict value of current position

Previous slide.

At the end of this semester, you will be able to understand the algorithms and network structure used to achieve these astonishing performances. Important are two types of outputs.

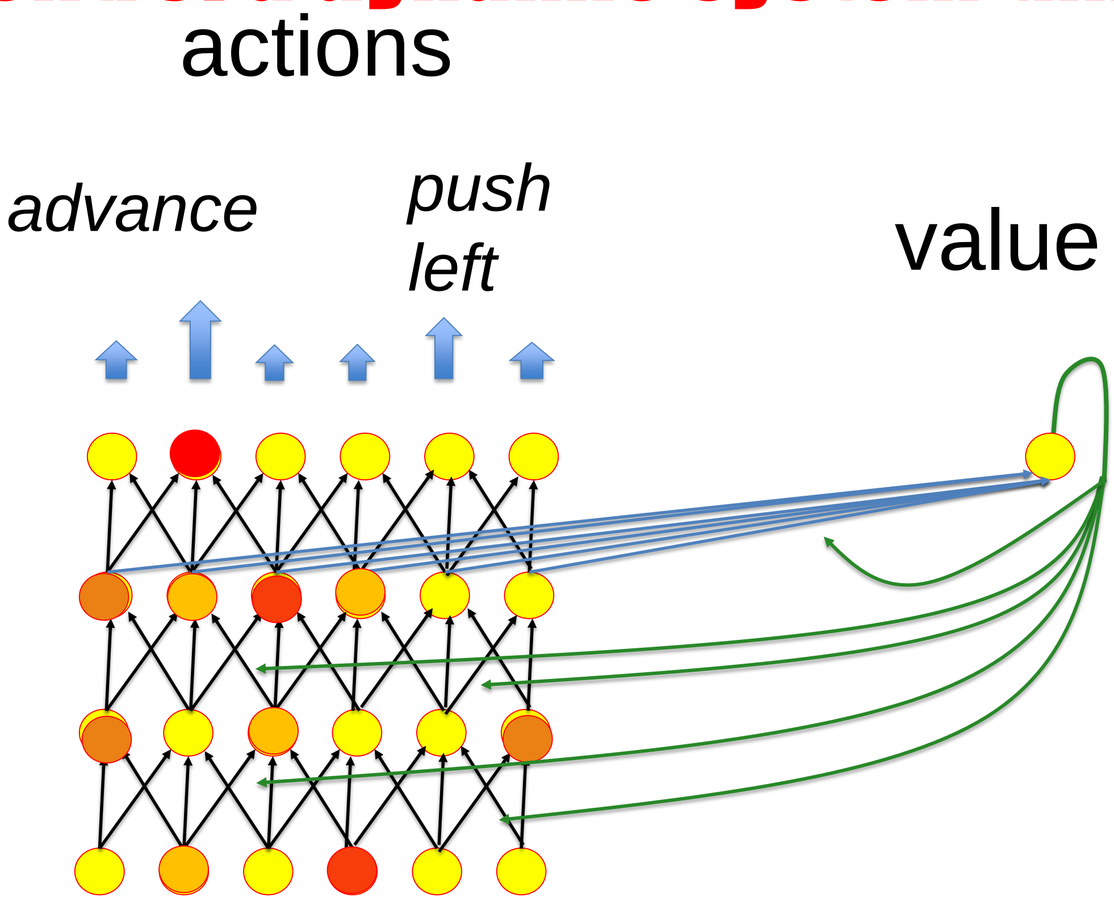
Left: different output neurons represent different actions.
Right: an additional output neuron represents the value of the present state; we can loosely define the value as the probability to win, or the ‘average reward’ that you can get starting from this state.

The input is a representation of the present state of the game.

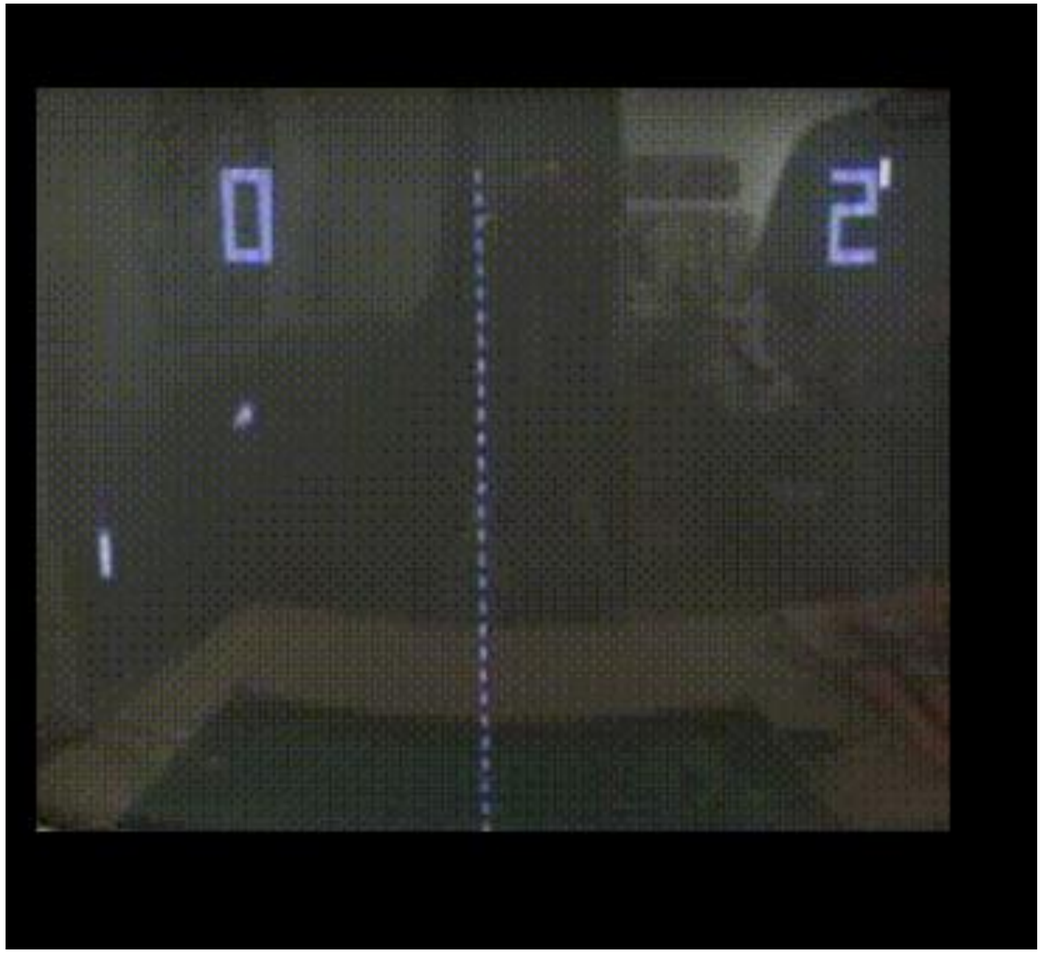
Details will become clear toward the end of the semester; at the moment the aim is just to give you a flavor of the high-level concepts.

Deep Reinforcement Learning:

Control a dynamic system (minprojects)



Example: Play Pong (Atari game)



Previous slide.

In the miniproject on RL, you will train a game. Training will be based on reward: successful behavior of the simulated agent will give positive rewards.

Quiz: Rewards in Reinforcement Learning

[ ] Reinforcement learning is based on rewards
[ ] Reinforcement learning aims at optimal action choices
[ ] In chess, the player gets an external reward after every move
[ ] In table tennis, the player gets a reward when he makes a point
[ ] A dog can learn to do tricks if you give it rewards at appropriate moments

Previous slide. Your notes.

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 2: Elements of Reinforcement Learning

- Examples of Reward-based Learning
- **Elements of Reinforcement Learning**

Previous slide.
We now start with the formalization of reinforcement learning

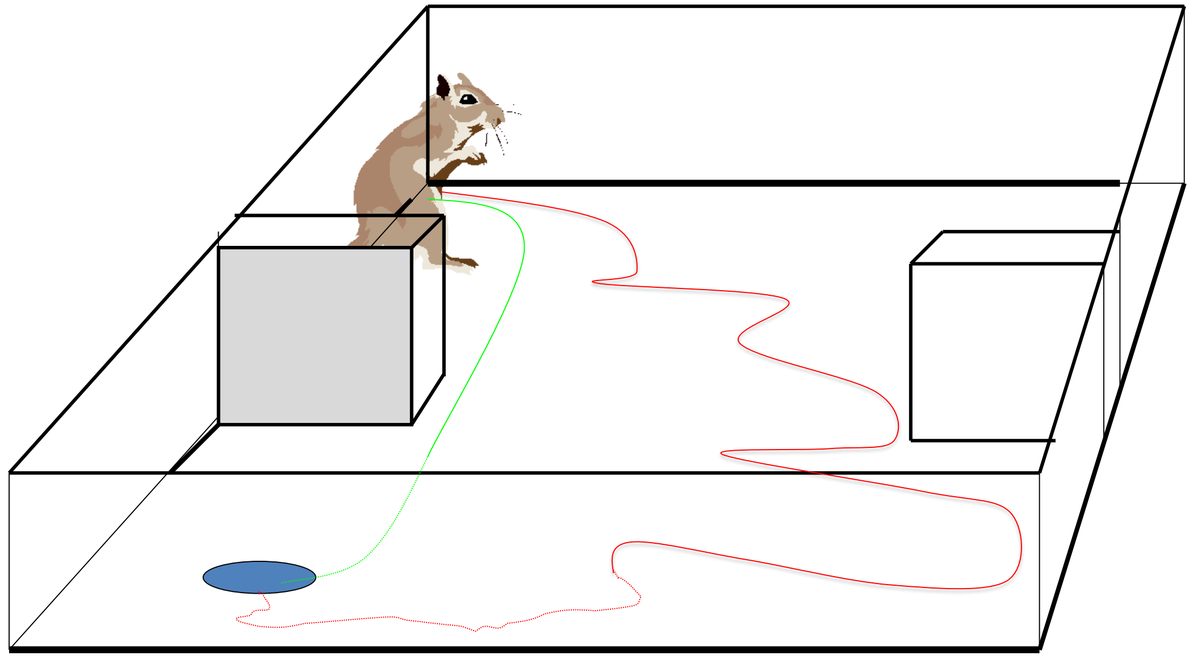
Elements of Reinforcement Learning:

-states
-actions
-rewards

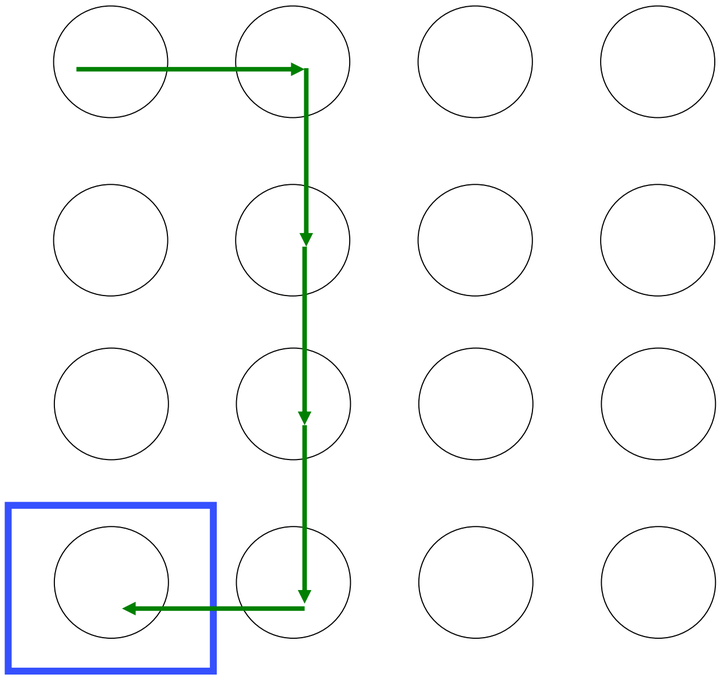
Previous slide.

Reinforcement learning needs states, actions, and rewards.

Elements of Reinforcement Learning:



- discrete states
- discrete actions
- sparse rewards



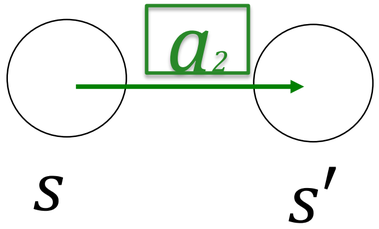
Previous slide.

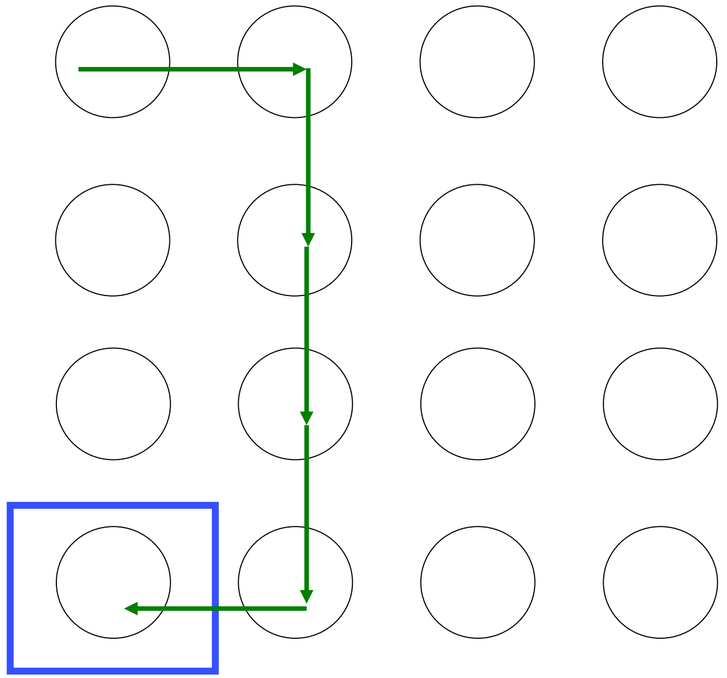
Note that, for standard formulations of Reinforcement Learning Theories this (normally) implies discretizing space and actions.

We will study continuous-space formulations only next week.

Elements of Reinforcement Learning:

- discrete states:
  - old state s
  - new state s'
- current state: s_t
- discrete actions: $a_1,\ a_2\ \ldots\ a_A$
- current action: a_t
- current reward: r_t





- Mean rewards for transitions:

$$R^{a}_{s \to s'}$$

often most transitions have zero reward

Previous slide.

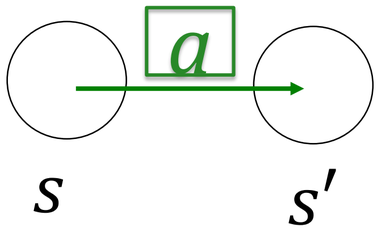
The elementary step is:
The agent starts in state s.
It takes action a
It arrives in a new state s’
Potentially receiving reward r (during the transition or upon arrival at s’).

Since rewards are stochastic we have to distinguish the mean reward at the transition (capital R with indices identifying the transition) from the actual reward (lower-case r with index t) that is received at time t on a transition.

Note that in many practical situations most transitions or states have zero rewards, except a single ‘goal’ state at the end.

States in Reinforcement Learning:

- discrete states:
  - starting state s
  - arrival state s'
- current state: s_t



state = current configuration/well-defined situation
= generalized ‘location’ of actor in environment

Previous slide.

What are these discrete states?
Loosely speaking a state is the current configuration that **uniquely** describes the momentary situation. We can think of the generalized ‘location’ of the actor in the environment

To get acquainted with this, let us look at an example.

Reinforcement Learning: Example Acrobot

3 actions: a_1 = no torque,

a_2 = torque +1 at elbow,

a_3 = torque -1 at elbow

States?

→ discretize!

reward if tip above line

Suppose 5 states per dimension, How many states in total?

[ ] 5

[ ] 25

[ ] 125

[ ] 625

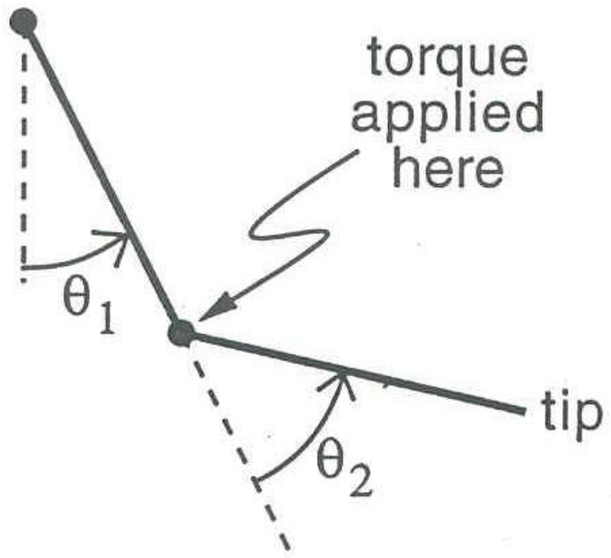


Figure 11.4 The acrobot.

From Book: Sutton and Barto

Previous slide.
The aim of the acrobat is to move the tip above the blue line. To achieve this torque can be applied at the ‘elbow’ link. The second link is the ‘shoulder’.

There are three possible actions.
But what are the states? How many states do we have?

Reinforcement Learning: Example Acrobot

1st episode: long sequence of random actions
400th episode: short sequence of ‘smart’ actions

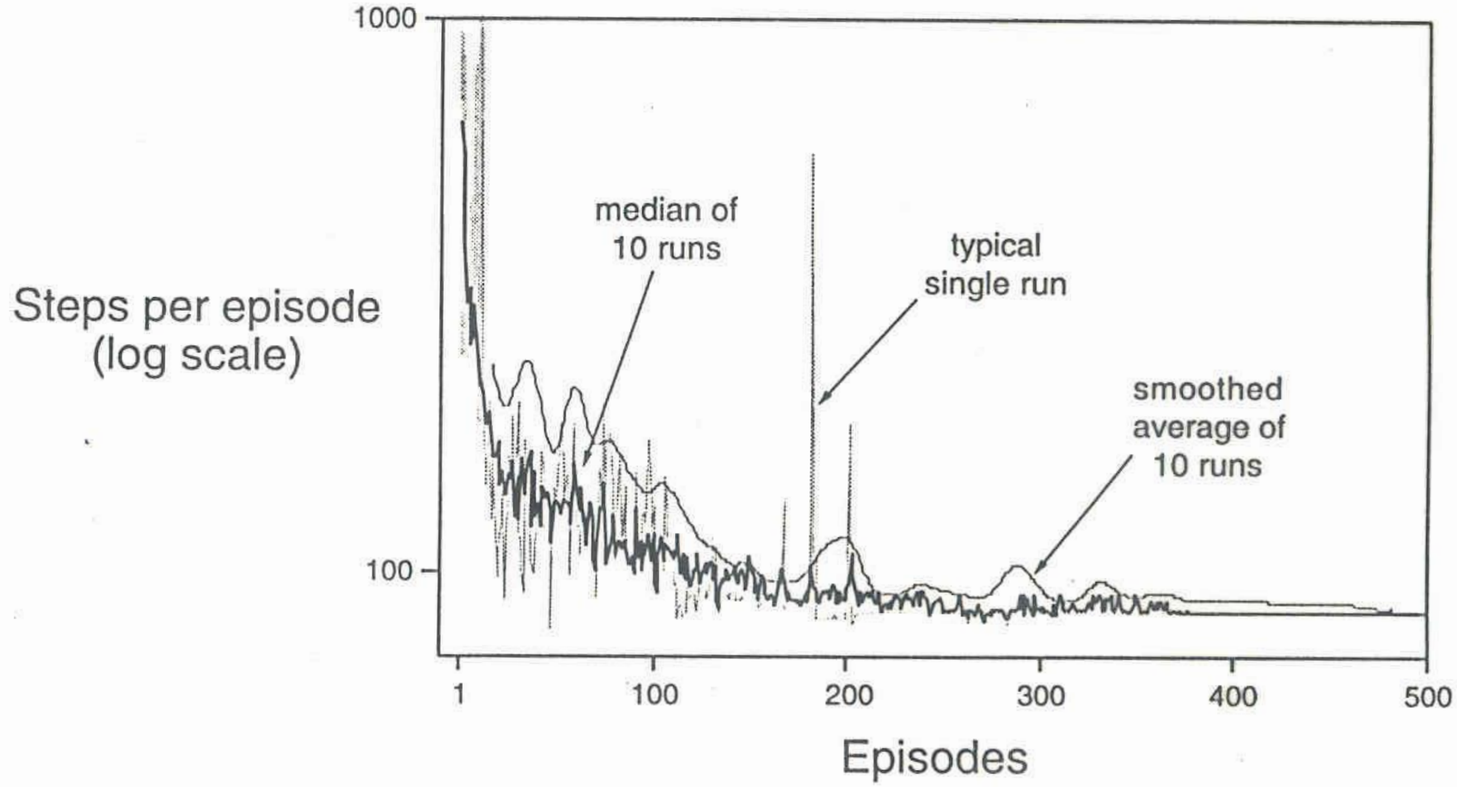


Figure 11.6 Learning curves for Sarsa(λ) on the acrobot task.

From Book:
Sutton and Barto

Previous slide.

An episode finishes if the target is reached. Over time episodes get shorter and shorter indicating that the acrobat has discovered (via reinforcement learning) a smart sequence of actions so as to reach the target (i.e., move the tip above the reference line)

Reinforcement Learning: Example Acrobot

after 400 episodes

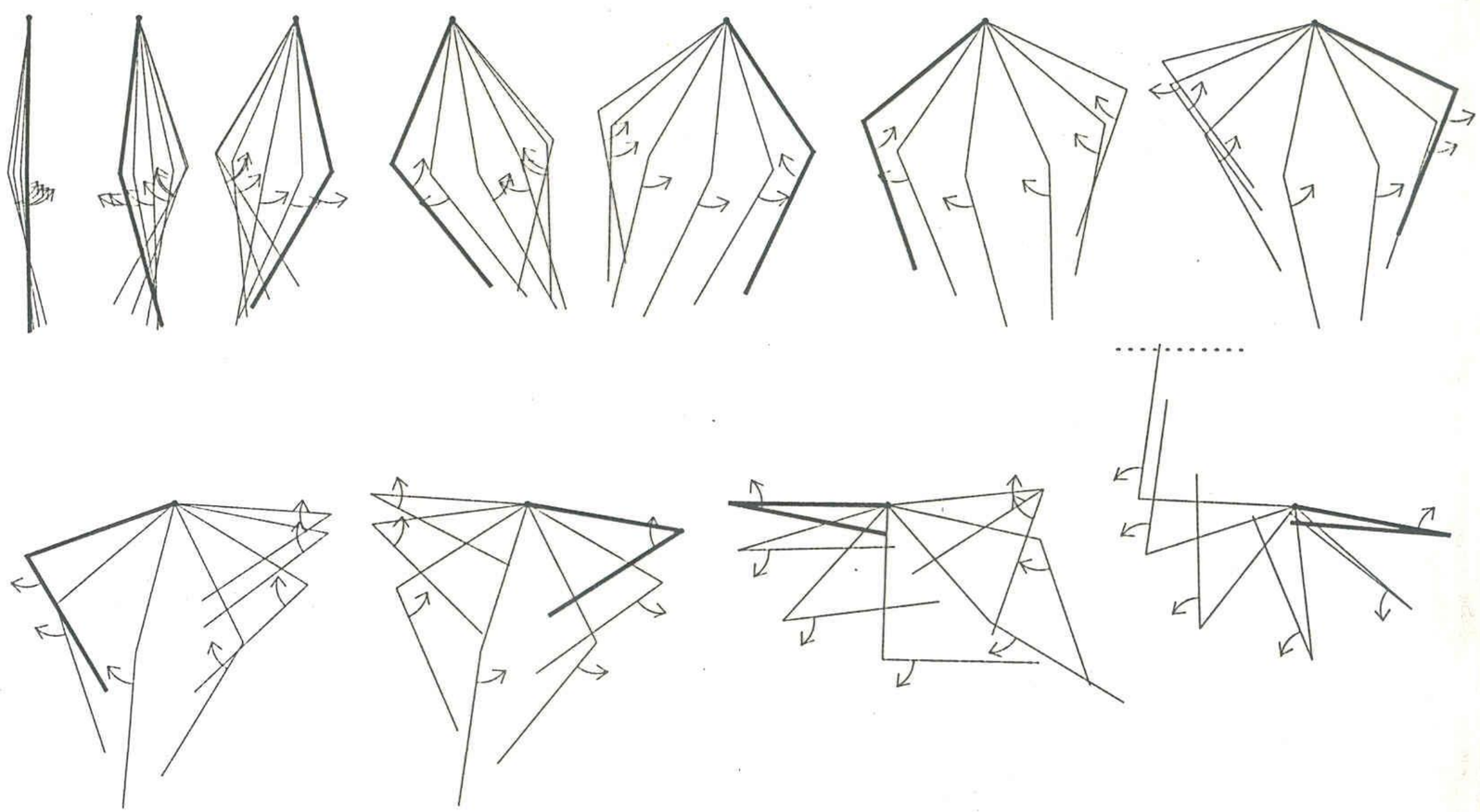


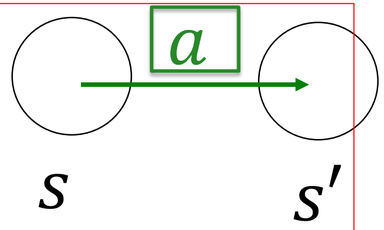
Figure 11.7 A typical learned behavior of the acrobot. Each group is a series of consecutive positions, the thicker line being the first. The arrow indicates the torque applied at the second joint.

From Book:
Sutton and Barto

Previous slide.
One example of an action sequence, after learning, is shown.

Summary: Elements of Reinforcement Learning

There can be MANY states
Often need to discretize first
(→ later we try to model in continuum)



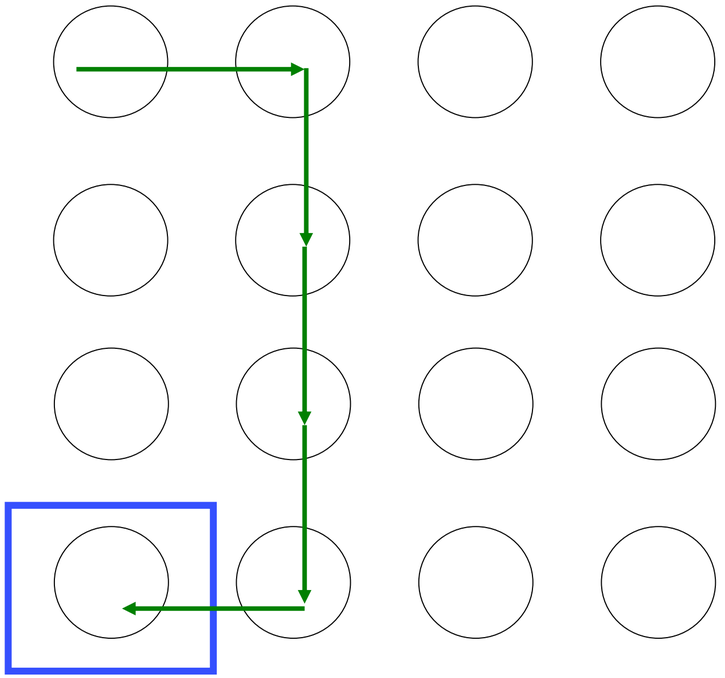
- discrete actions: a

- Mean reward for transition:

$$R^{a}_{s\to s'} = E\left(r\middle|s,a,s'\right)$$

- current actual reward: r_t

often most transitions have zero reward



Previous slide.
Conclusion: In all practical situations, there is an enormous number of states.

In many situations we can think of the actions as discrete. For the moment we also think of the states as discrete (but next week we will go to continuous state space)

Quiz: Reinforcement Learning for backgammon

Case Studies

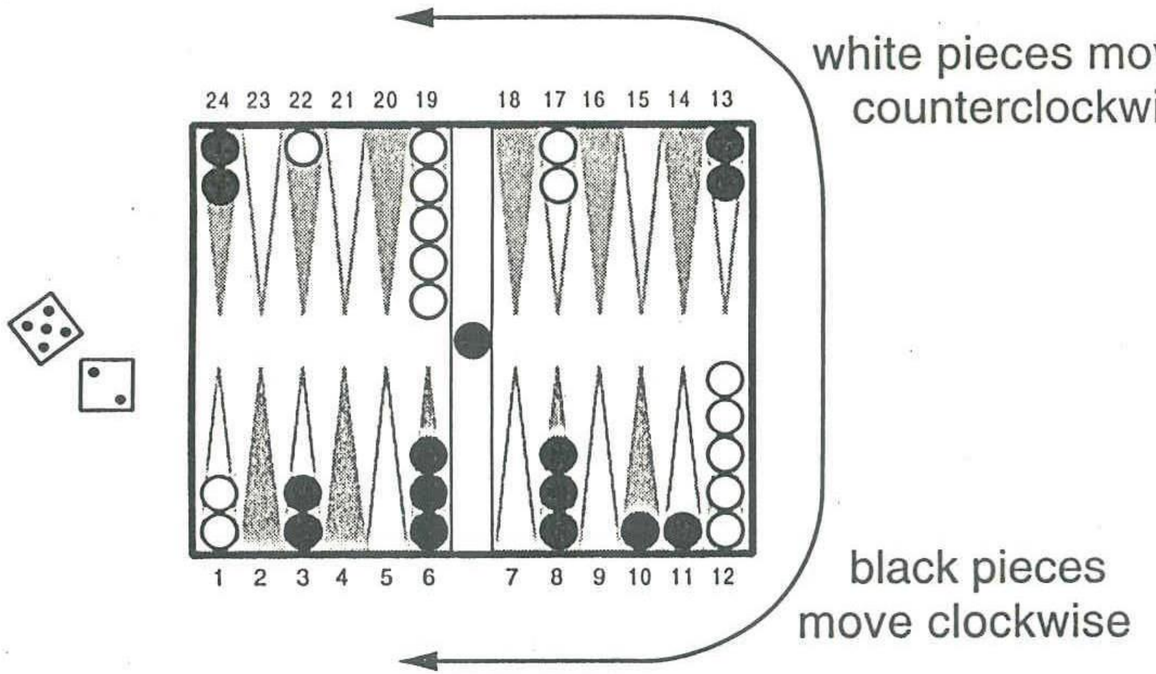


Figure 11.1 A backgammon position.

Game position = discrete states!

Suppose 2 pieces per player, How many states in total?

[ ] 100<n<500

[ ] 500<n<5000

[ ] 5 000<n<50 000

[ ] n>50 000

From Book:
Sutton and Barto

Previous slide.

Backgammon game. There are 24 fields on the board. Players have several pieces. Pieces are protected if there are two of the same color on the same field.

To make it simply, we now consider that both players have two pieces each left. How many different states are there in total?

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

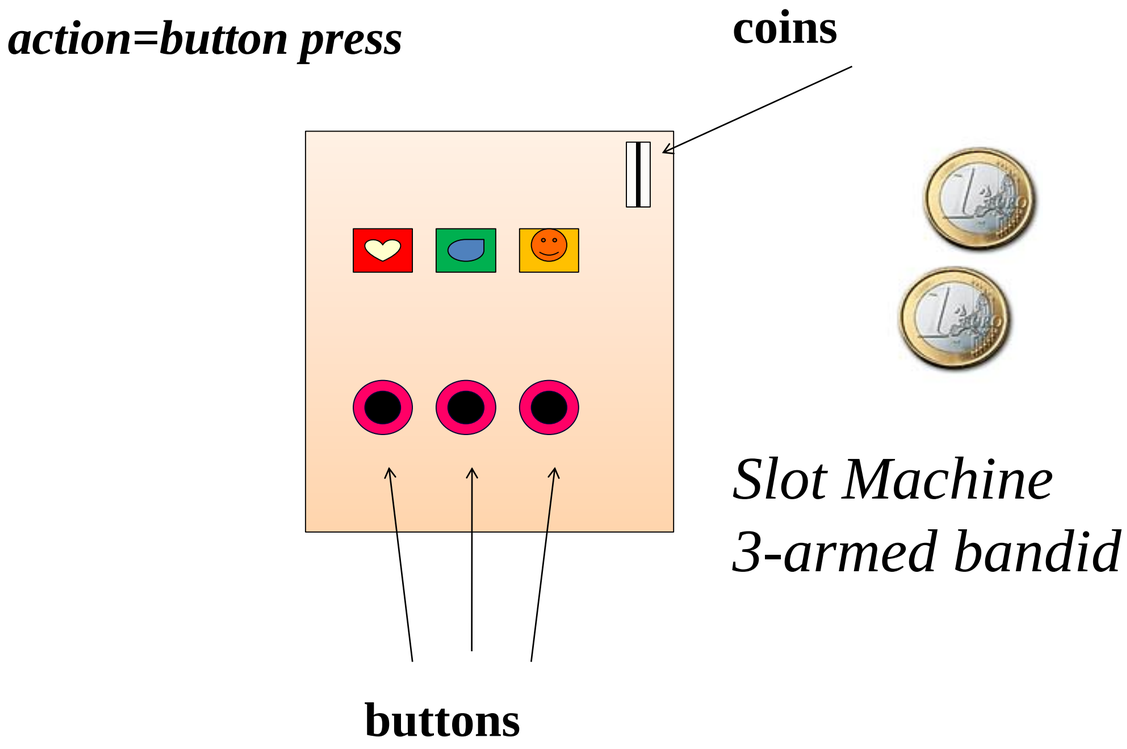
Part 3: One-step horizon (bandit problems)

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- **One-step horizon (bandit problems)**

Previous slide.

We start with the simplest discrete example: the game is over and reward is given after a single step.

One-step horizon games (bandit)

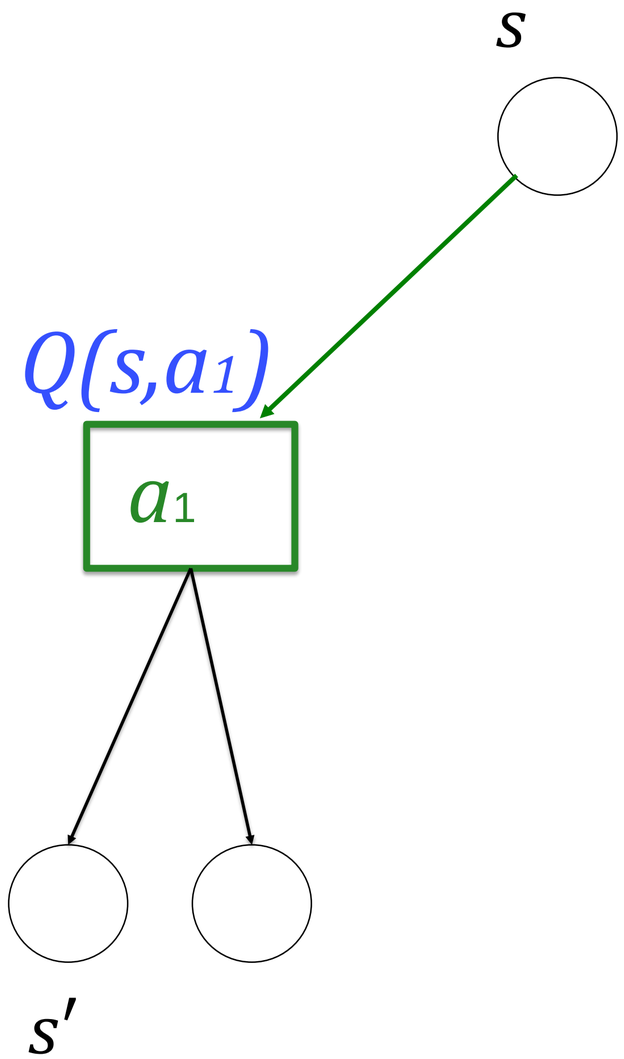


Slot Machine
3-armed bandid

Previous slide.
The standard example is a multi-armed bandit, or slot machine: you have to choose between a few actions, and once you have pressed the button you can just wait and see whether you get reward or not.

One-step horizon games

Q-value: $Q(s,a)$
Expected reward for action a starting from s



Blackboard1:
Q-values

Previous slide.

One of the most central notion in reinforcement learning is the Q-value.
Q(s,a) has two indices: you start in state s and take action a.
The Q-value Q(s,a) is (an estimate of) the mean expected reward that you will get if you take action a starting from state s.

One-step horizon games

Blackboard1:
Q-values

Your notes.

One-step horizon games: Q-value

Q-value $Q(s,a)$

Expected reward for action a starting from s

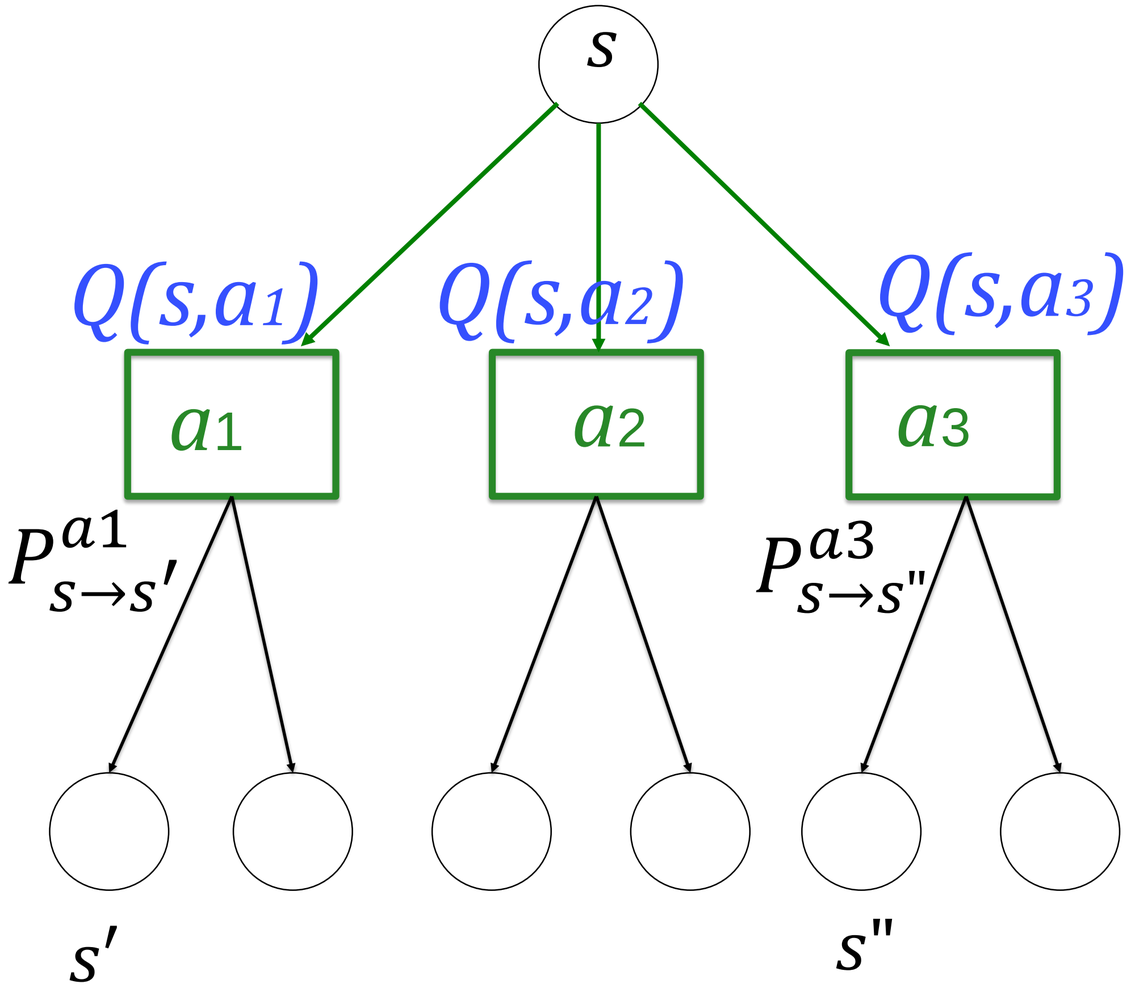
$$Q(s,a) = \sum_{s'} P^{a}_{s\to s'} R^{a}_{s\to s'}$$

Reminder:

$$R^{a}_{s\to s'} = E\left(r \middle| s', a, s\right)$$

Similarly:

$$Q(s,a) = E(r|s,a)$$



Now we know the Q-values: which action should you choose?

Previous slide.

$P_{s \to s'}^{a1}$ is the probability that you end up in a specific state s' if you take action a1 in state s.
We refer to this sometimes as the 'branching ratio' below the 'actions'.

Q(s,a) is attached to the branches linking the state s with the actions.

actions are indicated by green boxes; states are indicated by black circles.

The mean reward $R_{s \to s'}^{a}$ is defined as the expected reward given that you start in state s with action a and end up in state s' (see Blackboard 1).

Given the branching ratio and the mean rewards, it is easy to calculate the Q-values (Blackboard 1).

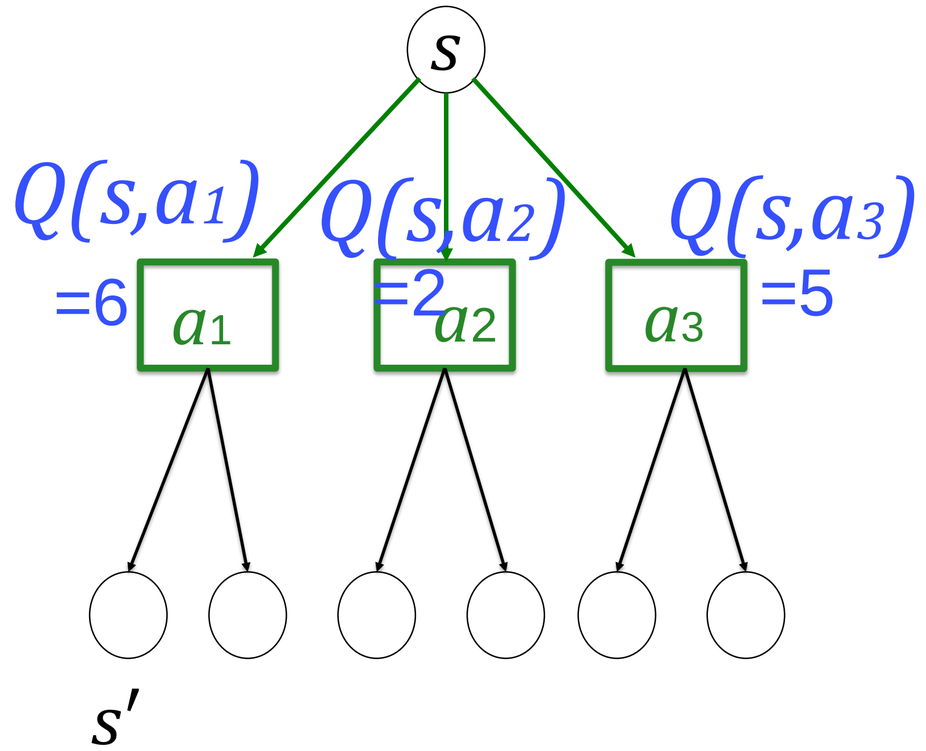
Optimal policy (greedy)

Suppose all Q-values are known:

take action a with*

$$Q(s,a^*) \geq Q(s,a_j)$$

other actions



optimal action:

$a^* = \text{argmax}_a\, [Q(s,a)]$

Optimal policy is also called ‘greedy policy’

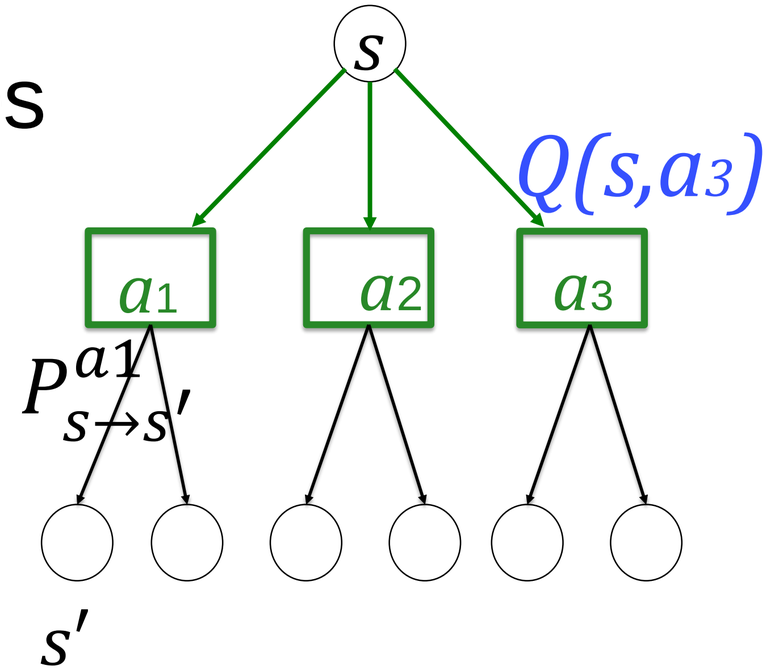
Previous slide.
And once you have the Q-values it is easy to choose the optimal action:
Just take the one with maximal Q-value.

One-step horizon games

Q-value = expected reward for state-action pair

If Q-value is known, choice of action is simple
→ take action with highest Q-value

BUT: we normally do not know the Q-values
→ estimate by trial and error



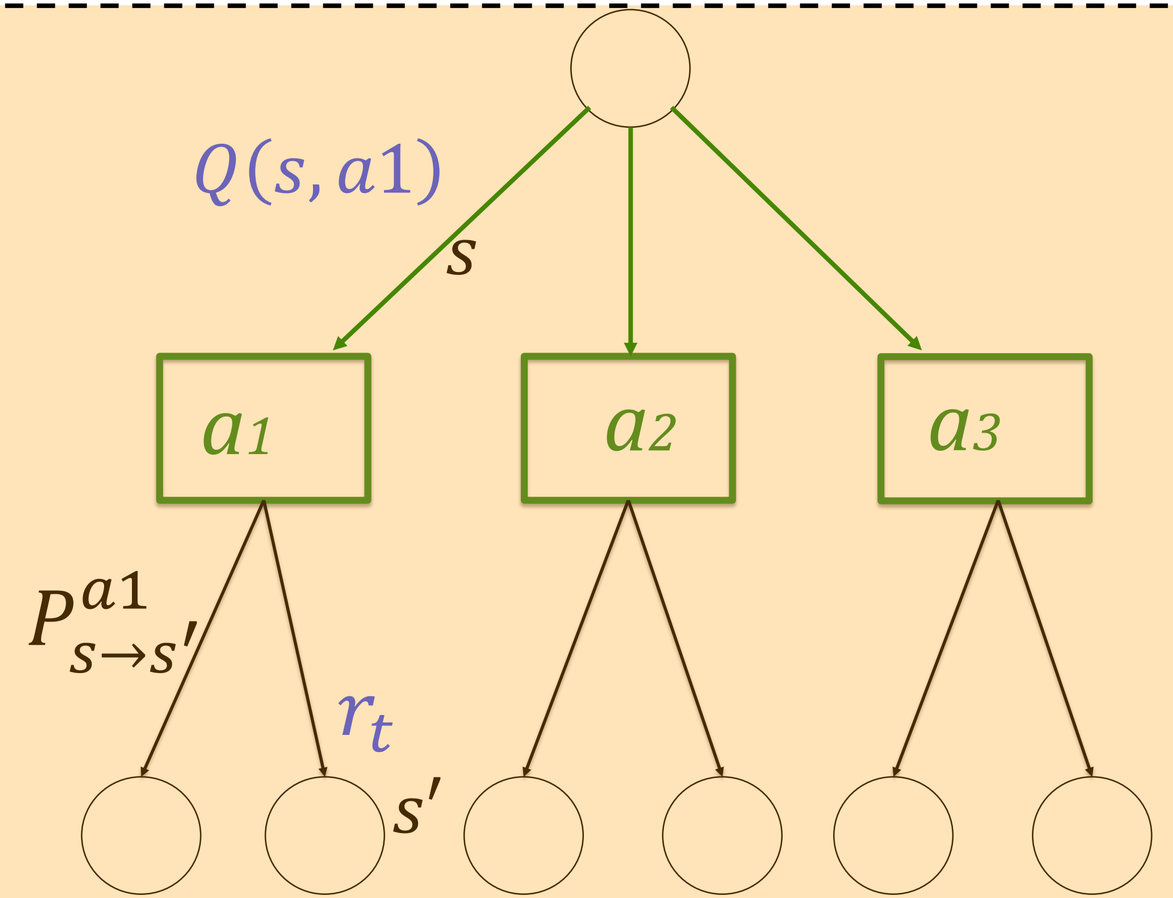
Previous slide.
The only remaining problem is that we do not know the Q-values, because the casino gives you neither the branching ratio nor the reward scheme.

Hence the only way to find out is by trial and error (that is, by playing many times – the casino will love this!).

Exercise 1 (from earlier session today)

Expected reward

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



Show that empirical averaging over k trials gives an update rule

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

Exercise 1 (in class)

Exercise 1. Iterative update (in class)

We consider an empirical evaluation of $Q(s, a)$ by averaging the rewards for action a over the first k trials:

$$Q_k = \frac{1}{k} \sum_{i=1}^{k} r_i.$$

We now include an additional trial and average over all $k + 1$ trials.

a. Show that this procedure leads to an iterative update rule of the form

$$\Delta Q_k = \eta(r_k - Q_{k-1}),$$

(assuming $Q_0 = 0$).

b. What is the value of η?

c. Give an intuitive explanation of the update rule. *Hint: Think of the following: If the actual reward is larger then my estimate, then I should ...*

Blackboard2: Exercise 1

Your notes.

Convergence in Expectation

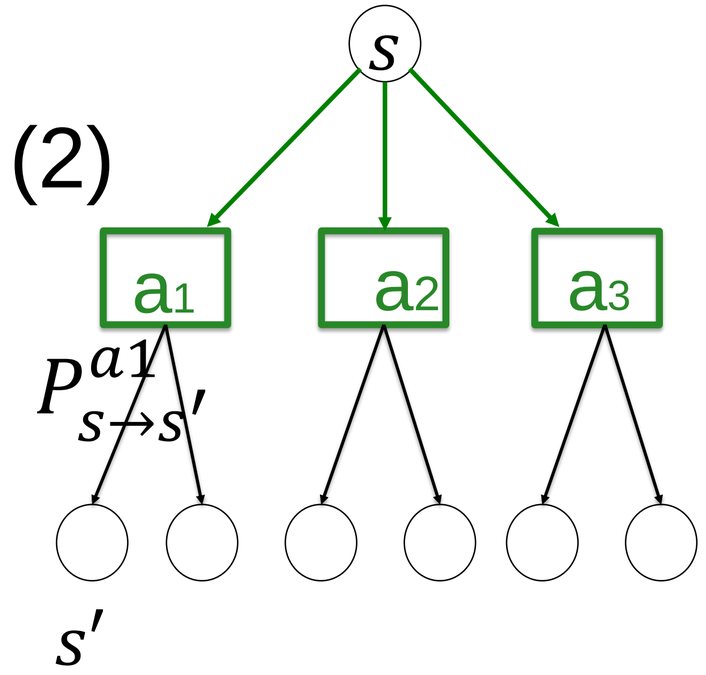
Proof of (i) will come: Blackboard3

After taking action a in state s, we update with

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)] \qquad (1)$$

(i) If (1) has **converged in expectation given (s,a),** then $\hat{Q}(s,a)$ has an expectation value,

$$E\left[\hat{Q}(s,a)\right] = \sum_{s'} P^{a}_{s\to s'} R^{a}_{s\to s'} = Q(s,a) \qquad (2)$$



(ii) If the learning rate η decreases, fluctuations around the **empirical mean** $\langle\hat{Q}(s,a)\rangle$ decrease and **the empirical mean approaches** $Q(s,a)$

Previous slide.
When evaluating the **expectation value given (s,a),** the learning rate drops out since we set the left-hand-side to zero. The exact value of η is not relevant, as discussed in the theorem. Part (i) of the theorem states that the expectation value of $\hat{Q}(s,a)$ is the correct Q-value. For a quick proof of part (i) see the video. On the blackboard a stronger statement was shown.
Convergence in expectation is equivalent to imagining that you start millions of trials with the same value $\hat{Q}(s,a)$ without any intermediate update. So in that sense it is like a super-big 'batch' of examples.
In practice, we do not have expectations but online updates with fluctuations. It is important is that η is small at the end of learning so as to limit the amount of fluctuations.
Part (ii) states that **online mean** for small learning rate also goes to the correct Q-value. Indeed, since the equations are linear (for the bandit problem = 1-step horizon), the calculation of part (i) apply analogously to the long-term empirical temporal average (denoted by angular brackets)
$\langle \Delta \hat{Q}(s,a) \rangle = \eta \langle [r_t - \hat{Q}(s,a)] \rangle$

This equivalence based on linearity is not true for the multi-step horizon that we discuss later in this lecture.

Proof: Convergence in Expectation

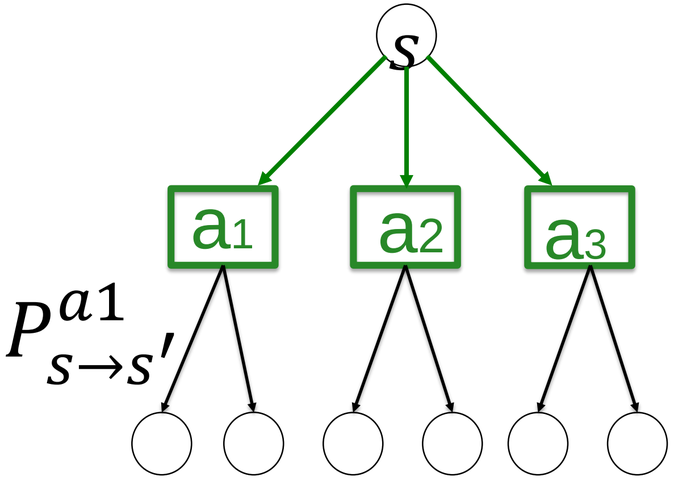
Blackboard3: Proof of (i)

After taking action a in state s, we update with

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)\,] \qquad (1)$$

(i) If (1) has converged in expectation, then $\hat{Q}(s,a)$ has an expectation value,

$$E\left[\widehat{Q}(s,a)\right] = \widehat{Q}(s,a) = \sum_{s'} P^{a}_{s\to s'}\, R^{a}_{s\to s'} = Q(s,a) \qquad (2)$$



Your notes.

Blackboard3

Part (i) of Theorem
converged in expectation → $E(\Delta\hat{Q}(s,a)|s,a)=0$

expectation of all possible futures with correct statistical weight

we always start in (s,a) while the system is frozen

Perspective similar to a batch mode:
update only **after** (infinitely) many trials that all start in (s,a) with the same value $\hat{Q}(s,a)$
=
update the expectation over all possibilities that may occur in the next time step.

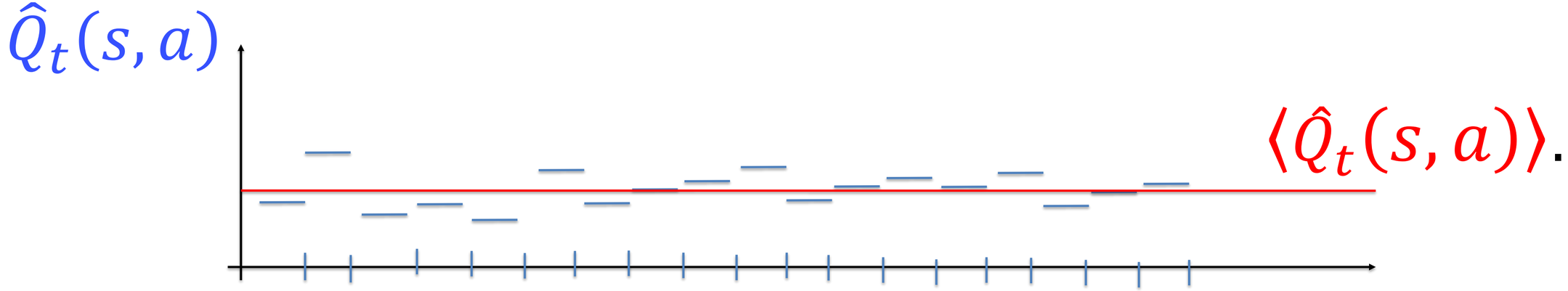
Previous slide:
$\hat{Q}(s,a)$ denotes the current estimate of the Q-value. Claim: If Q no longer changes (in expectation) then it must be the correct Q-value.

There are different views on how to interpret the ‘expectation;:
- Formally from a mathematical point of view: average over all possible outcomes of the next time step given (s,a).
- In a simulation this would correspond to the following sampling procedure: You freeze the value of $\hat{Q}(s,a)$ and run MANY times (N to infinity) a test with the state-action pair (s,a) as a starting condition. Then you evaluate the resulting ‘batch update’ averaged across all these examples. If the batch update with Millions of Examples is zero, that implies that you have converged to the correct value.

Blackboard3

Part (ii) of Theorem:

We work with the online update $\Delta\hat{Q}(s,a)$. With finite learning rate, the value of $\hat{Q}_t(s,a)$ fluctuates around a mean$\langle\hat{Q}_t(s,a)\rangle$.



Under the hypothesis of the theorem, the mean is equal to the ‘correct’ Q-value./

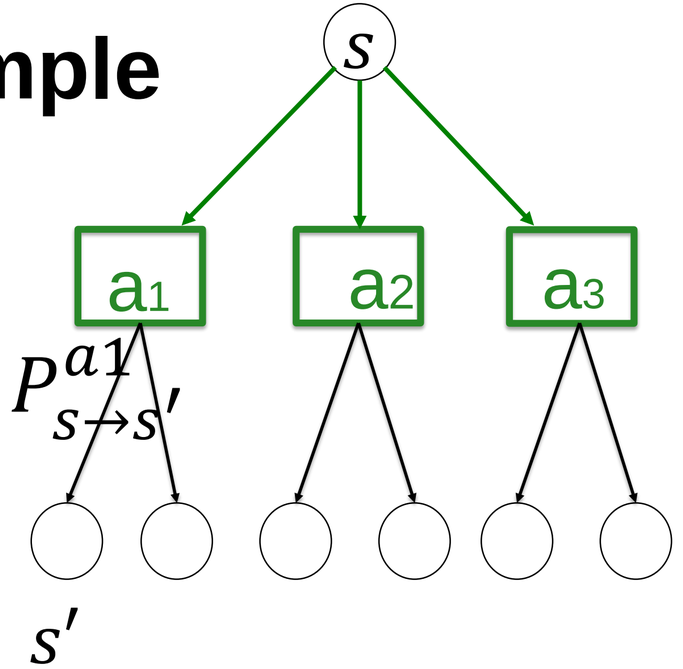
Your notes. (Proof in the Blackboard notes)

One-step horizon: summary

Q-value = expected reward for state-action pair

If Q-value is known, choice of action is simple

→ take action with highest Q-value



If Q-value not known:

→ estimate $\hat{Q}$ by trial and error

→ update with rule

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)\,] \qquad \textbf{(1)}$$

→Let learning rate η decrease over time

Iterative algorithm (1) converges in expectation

Previous slide.
Let us distinguish the ESTIMATE $\hat{Q}(s,a)$ from the real Q-value $Q(s,a)$

The update rule can be interpreted as follows:
if the actual reward is larger than (my estimate of) the expected reward, then I should increase (a little bit) my expectations.

The learning rate η :

In exercise 1, we found a rather specific scheme for how to reduce the learning rate over time. But many other schemes also work in practice. For example you keep η constant for a block of time, and then you decrease it for the next block.

Note: in later lectures I will often use the symbol α instead of η

Both symbols indicate what is called the ‘learning rate’ in Deep Learning.

Artificial Neural Networks: RL1 (continued)

Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Parts 4-6: Examples of Reward-based Learning

Objectives for Lecture RL1:

- Reinforcement Learning (RL) is learning by rewards ✓
- Agents and actions, states and rewards ✓
- Convergence in expectation
- Exploration vs Exploitation
- Bellman equation
- SARSA algorithm

Reading:
Sutton and Barto, Reinforcement Learning (MIT Press, 2nd edition 2018)
Chapters: 1.1-1.4; 2.1-2.6; 3.1-3.5; 6.4

Reading for this week:

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

Chapters: 1.1-1.4; 2.1-2.6; 3.1-3.5; 6.4

Background reading:

Silver et al. 2017, Archive

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

Recall: Learning by reward

Learning by reward ‘’
BUT:
Reward is rare:
‘sparse feedback’ after
a long action sequence

Previous slide.

How does a human learn to play table tennis: How does a child learn to play the piano? How does a dog learn to perform tricks?

In all these cases there is no supervisor. No master guides the hand of the players during the learning phase. Rather the player ‘discovers’ good movements by rather coarse feedback. For example, the ball in table tennis does not land on the table as it should. That is bad (negative feedback). The ball has a great spin so that the opponent does not get. This is good (positive feedback).

Similarly, it is hard to tell a dog what to do. But if you reinforce the dog’s behavior by giving a ‘goodie’ at the moment when it spontaneously performs a nice action, then it can learn quite amazing things.

In all these cases it is the ‘reward’ that guides the learning. Rewards can be the goodie for the dog, or just the feeling ‘now I did well’ for humans.

Recall: Value is important

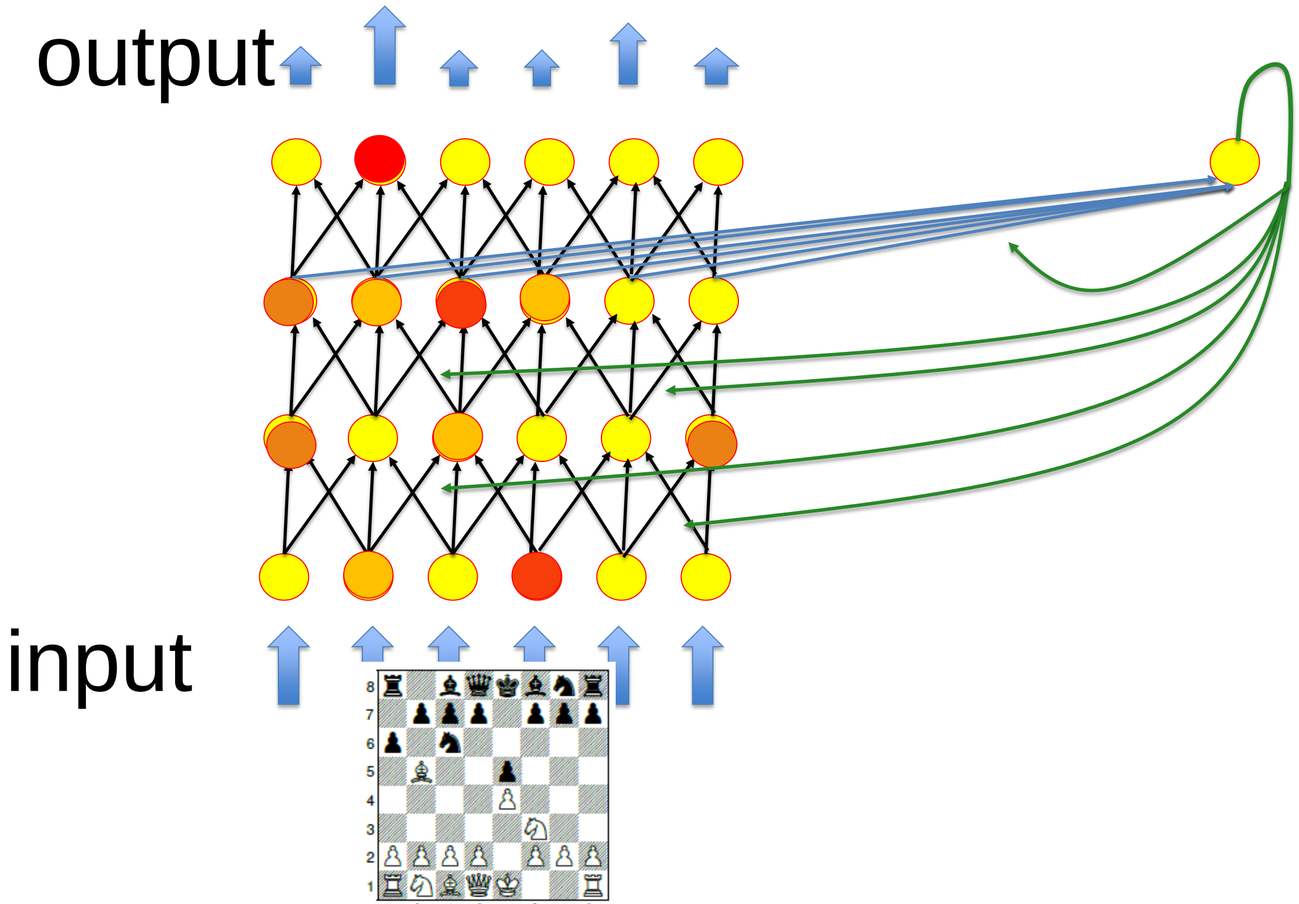
Network for choosing action

action:

Advance king

2nd output for **value** of state:

probability to win = expected reward



Learning by success signal

- change connections

aim:

- choose next action to win

aim for value unit:

- predict value of current position

Previous slide.

At the end of this semester, you will be able to understand the algorithms and network structure used to achieve these astonishing performances. Important are two types of outputs.

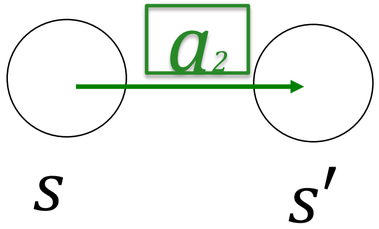
Left: different output neurons represent different actions.
Right: an additional output neuron represents the value of the present state; we can loosely define the value as the probability to win, or the ‘average reward’ that you can get starting from this state.

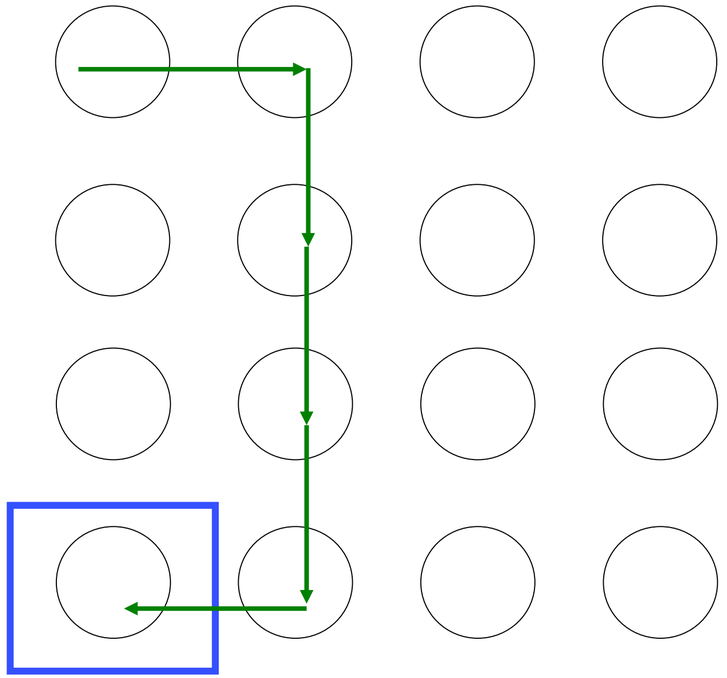
The input is a representation of the present state of the game.

Details will become clear toward the end of the semester; at the moment the aim is just to give you a flavor of the high-level concepts.

Recall: Elements of Reinforcement Learning:

- discrete states:
  - old state s
  - new state s'
- current state: s_t
- discrete actions: $a_1,\ a_2\ \ldots\ a_A$
- current action: a_t
- current reward: r_t





- Mean rewards for transitions:

$$R^{a}_{s \to s'}$$

often most transitions have zero reward

Previous slide.

The elementary step is:
The agent starts in state s.
It takes action a
It arrives in a new state s’
Potentially receiving reward r (during the transition or upon arrival at s’).

Since rewards are stochastic we have to distinguish the mean reward at the transition (capital R with indices identifying the transition) from the actual reward (lower-case r with index t) that is received at time t on a transition.

Note that in many practical situations most transitions or states have zero rewards, except a single ‘goal’ state at the end.

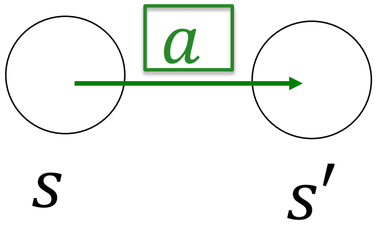
Recall: Elements of Reinforcement Learning

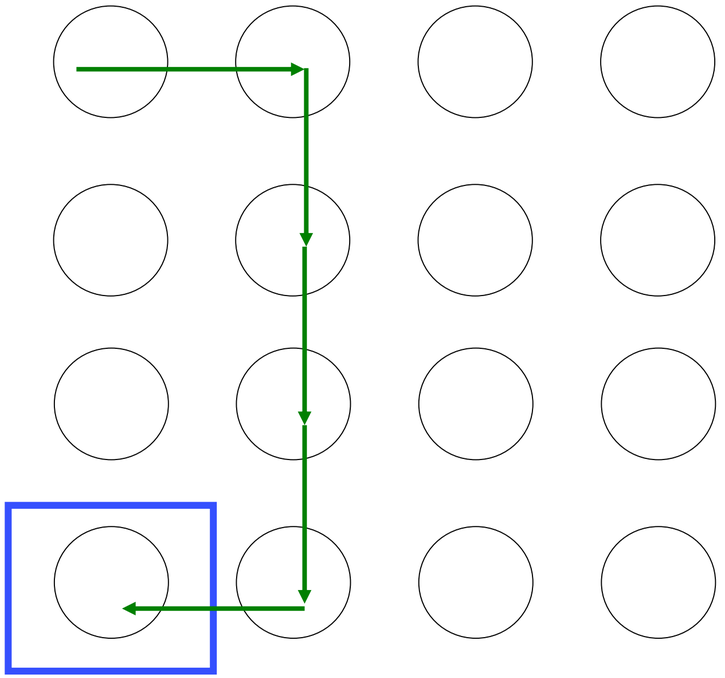
- Environment with discrete states
- Transitions (potentially stochastic) are driven by actions
- **Aim: choose good actions to optimize reward**
  → Markov Decision Model
  - Mean reward for transition:

$$R^{a}_{s \to s'} = E\left(r \middle| s, a, s'\right)$$

- current actual reward: r_t

often most transitions have zero reward





Previous slide.
Conclusion: In all practical situations, there is an enormous number of states.

In many situations we can think of the actions as discrete.
For the moment we also think of the states as discrete (but next week we will go to continuous state space).

Transitions between actions are influenced by action choices.

Transitions can be stochastic.

Actions should be chosen so as to maximize the reward. This setting is also known as Markov Decision Problem (MDP).

Recall: Q-value for one-step horizon games/bandit problem

Q-value $Q(s,a)$

Expected reward for action a starting from s

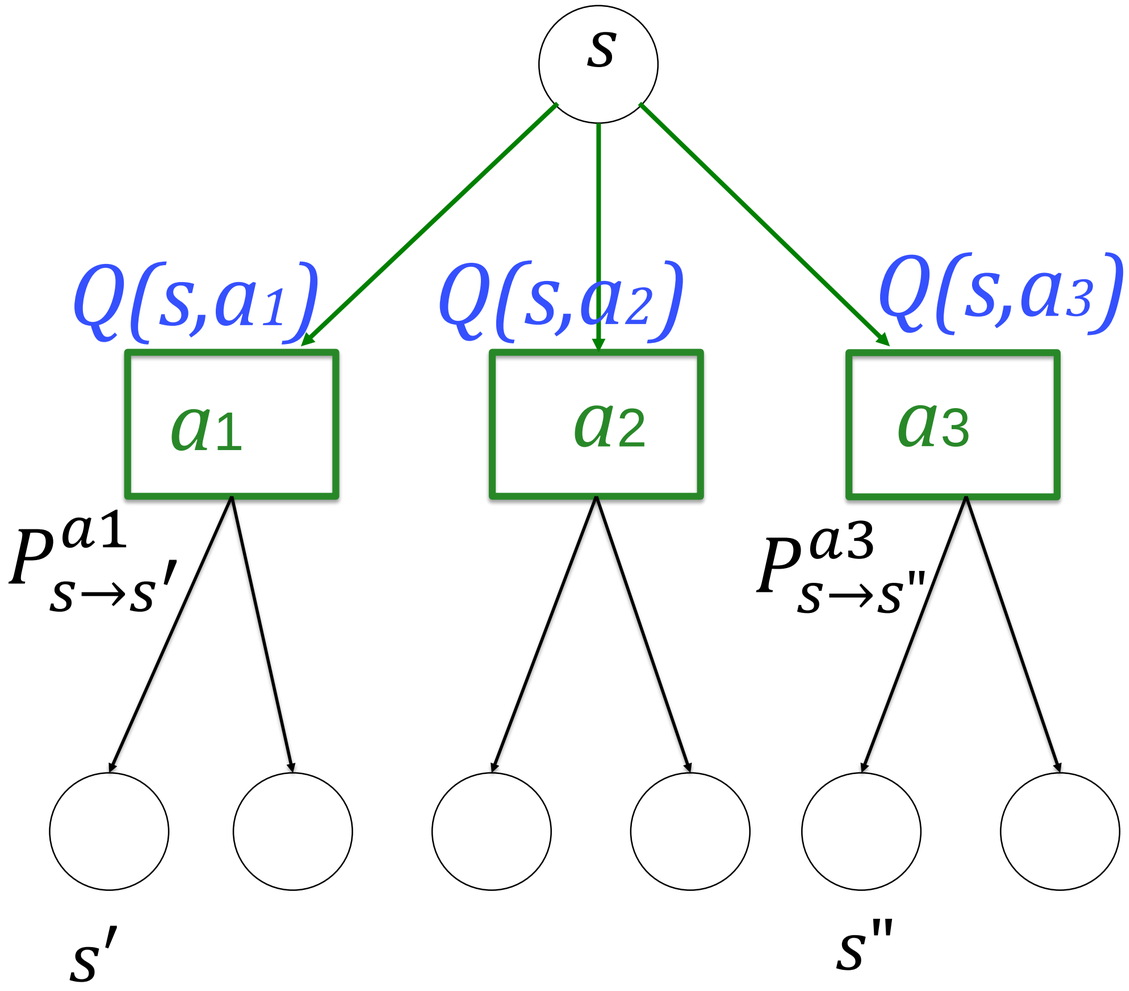
$$Q(s,a) = \sum_{s'} P^{a}_{s\to s'} R^{a}_{s\to s'}$$

Reminder:

$$R^{a}_{s\to s'} = E\left(r \middle| s', a, s\right)$$

Similarly:

$$Q(s,a) = E(r|s,a)$$



Now we know the Q-values: which action should you choose?

Previous slide.

$P_{s \to s'}^{a1}$ is the probability that you end up in a specific state s' if you take action a1 in state s.
We refer to this sometimes as the 'branching ratio' below the 'actions'.

Q(s,a) is attached to the branches linking the state s with the actions.

actions are indicated by green boxes; states are indicated by black circles.

The mean reward $R_{s \to s'}^{a}$ is defined as the expected reward given that you start in state s with action a and end up in state s' (see Blackboard 1).

Given the branching ratio and the mean rewards, it is easy to calculate the Q-values (Blackboard 1).

Recall: One-step horizon games (bandit problem)

Q-value = expected reward for state-action pair

If Q-value is known, choice of action is simple

→ take action with highest Q-value

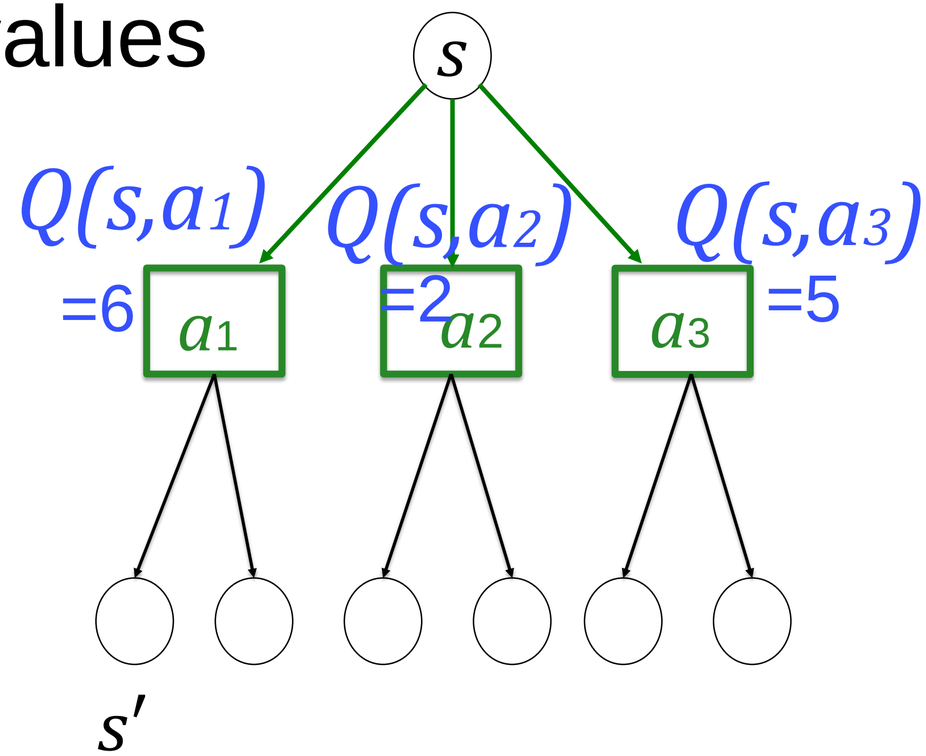
BUT: we normally do not know the Q-values

→ **estimate by online update**

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

Learning rate: often η , often α

$$\Delta Q(s,a) = \alpha\,[r_t - Q(s,a)]$$



Previous slide.
The only remaining problem is that we do not know the Q-values, because the casino gives you neither the branching ratio nor the reward scheme.

Hence the only way to find out is by trial and error (that is, by playing many times – the casino will love this!).

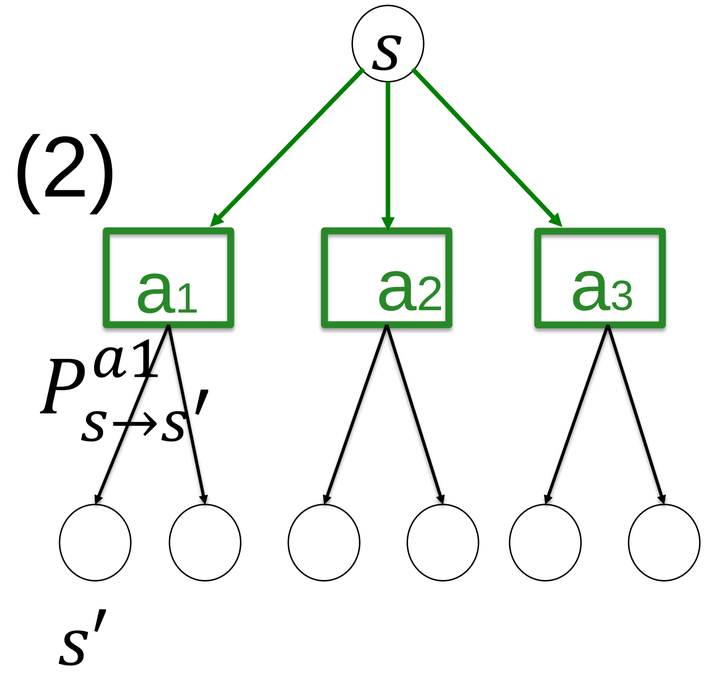
Recall: Update rule in Expectation (Theorem)

After taking action a in state s, we update with

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)] \qquad (1)$$

(i) If the expectation of the update rule (1) **given (s,a)** vanishes, then $\hat{Q}(s,a)$ has an expectation

$$E\left[\widehat{Q}(s,a)\right] = \sum_{s'} P^{a}_{s\to s'}\, R^{a}_{s\to s'} = Q(s,a) \qquad (2)$$



claim: 'I can even drop the expectation sign in (2)

(ii) If the learning rate η decreases, fluctuations around the **empirical mean** $\langle\widehat{\boldsymbol{Q}}(\boldsymbol{s},\boldsymbol{a})\rangle$ decrease and **the empirical mean approaches** $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$

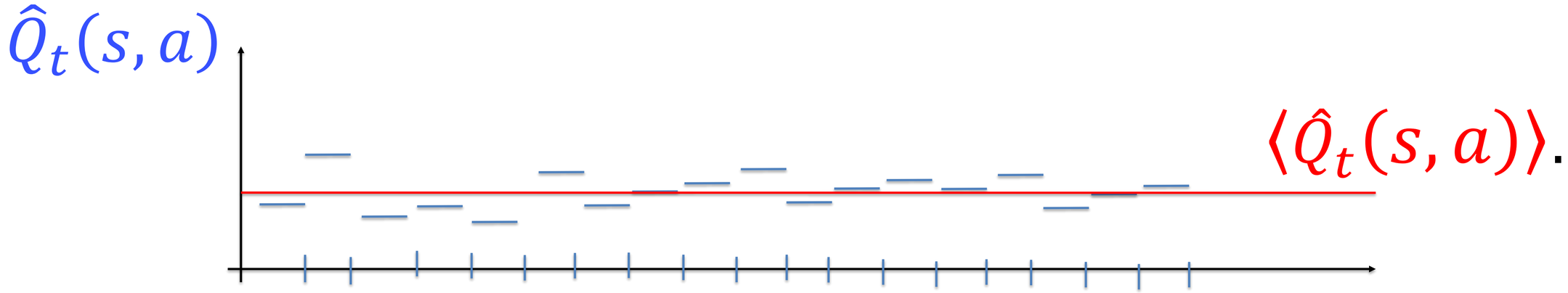
Previous slide.
When evaluating the **expectation value given (s,a),** the learning rate drops out since we set the left-hand-side to zero. The exact value of η is not relevant, as discussed in the theorem. Part (i) of the theorem states that the expectation value of $\hat{Q}(s,a)$ is the correct Q-value. For a quick proof of part (i) see the video. On the blackboard a stronger statement was shown. Today we make the statements more precise Convergence in expectation is equivalent to imagining that you start millions of trials with the same value $\hat{Q}(s,a)$ without any intermediate update. So in that sense it is like a super-big 'batch' of examples.
In practice, we do not have expectations but online updates with fluctuations. It is important is that η is small at the end of learning so as to limit the amount of fluctuations.
Part (ii) states that **online mean** for small learning rate also goes to the correct Q-value. Indeed, since the equations are linear (for the bandit problem = 1-step horizon), the calculation of part (i) apply analogously to the long-term empirical temporal average (denoted by angular brackets)

$$\langle \Delta \hat{Q}(s,a) \rangle = \eta \langle [r_t - \hat{Q}(s,a)] \rangle$$

This equivalence based on linearity is not true for the multi-step horizon that we discuss later in this lecture.

Part (ii) of Theorem:

We work with the **online update** $\Delta\hat{Q}(s,a)$. With finite learning rate, the value of $\hat{Q}_t(s,a)$ fluctuates around a mean $\langle\hat{Q}_t(s,a)\rangle$.



Claim: Under the hypothesis $\langle\Delta\hat{Q}(s,a) = 0\rangle$, the mean $\langle\hat{Q}_t(s,a)\rangle$ is equal to the 'correct' Q-value.

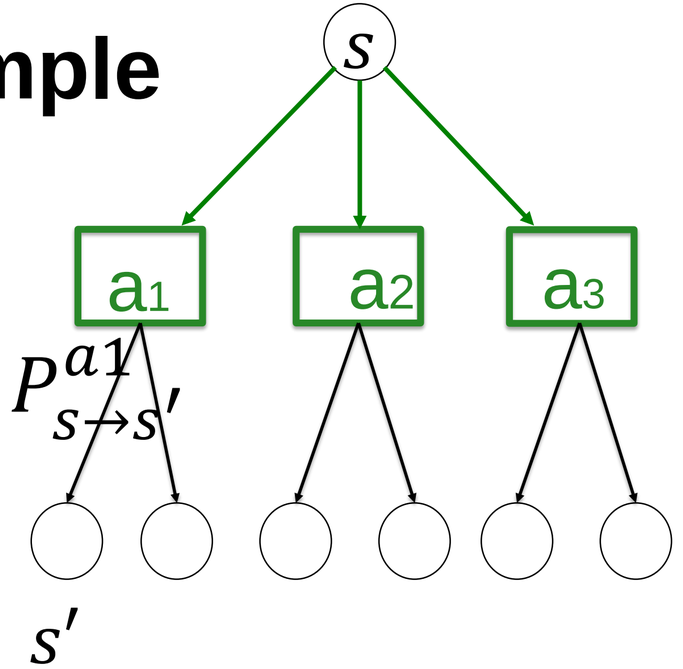
Your notes. (Proof in the Blackboard notes)

One-step horizon: summary

Q-value = expected reward for state-action pair

If Q-value is known, choice of action is simple

→ take action with highest Q-value



If Q-value not known:

→ estimate $\hat{Q}$ by trial and error

→ update with rule

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)\,] \qquad \textbf{(1)}$$

→Let learning rate η decrease over time

Iterative algorithm (1) fluctuates, but 'makes sense'

Previous slide.
Let us distinguish the ESTIMATE $\hat{Q}(s,a)$ from the real Q-value $Q(s,a)$

The update rule can be interpreted as follows:
if the actual reward is larger than (my estimate of) the expected reward, then I should increase (a little bit) my expectations.

The learning rate η :

In exercise 1, we found a rather specific scheme for how to reduce the learning rate over time. But many other schemes also work in practice. For example you keep η constant for a block of time, and then you decrease it for the next block.

Note: in later lectures I will often use the symbol α instead of η

Both symbols indicate what is called the ‘learning rate’ in Deep Learning.

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 3b: Expectation, Batch, and ONLINE rules

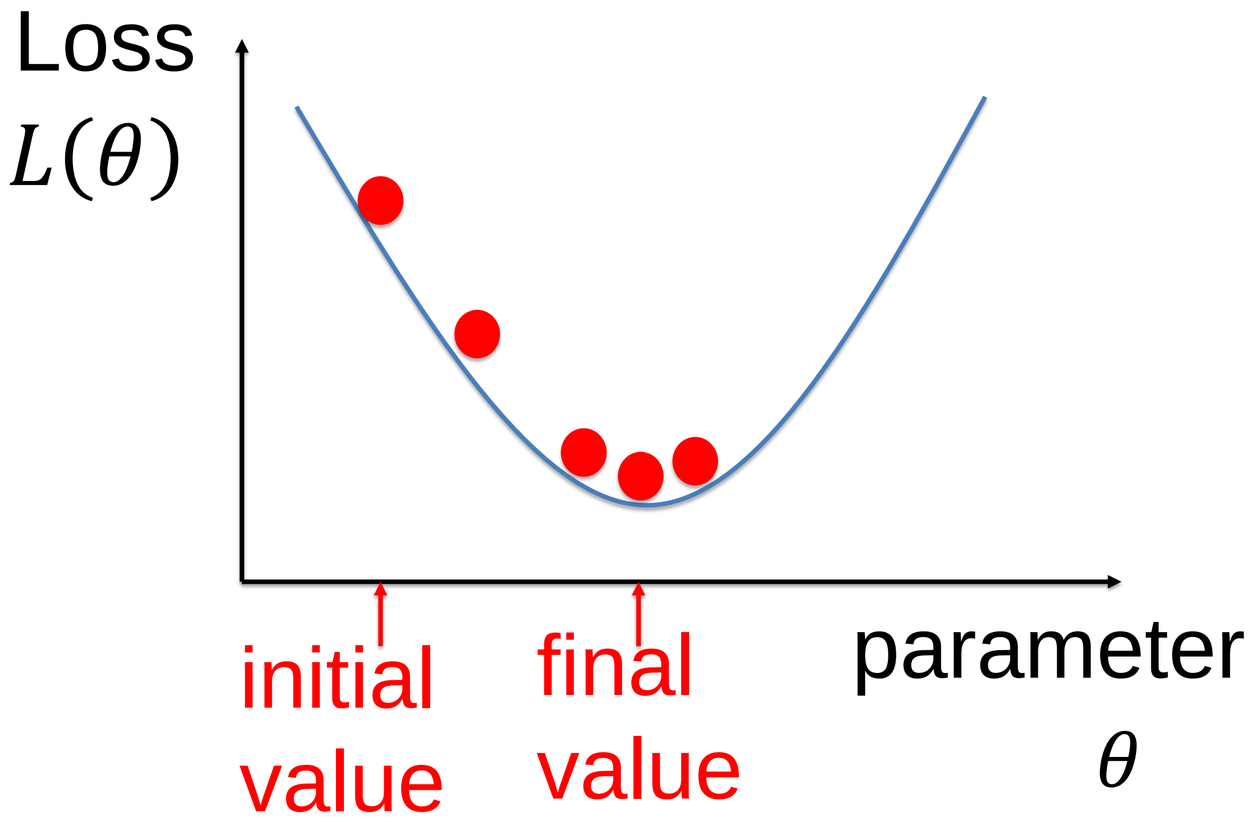
- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step horizon (bandit problems)
- **Expectation, batch, and online rules**

Previous slide.

Last week we did a first calculation with expectations and I argued that this is ‘batch-like’. Since the type of calculation is important, but since this comparison caused many questions after class, I add a detour.

All the material in this part is in principle standard material in Machine Learning classes, even though I put the accent on aspects that are important for Reinforcement Learning.

Detour Machine Learning ReCap: Online, Batch, Expectation



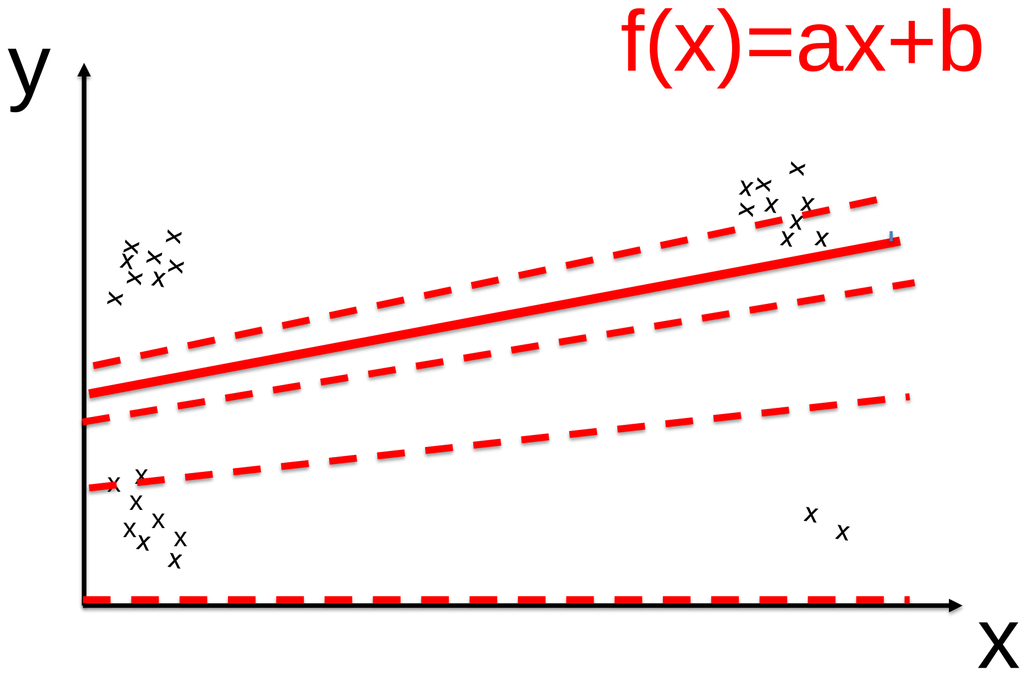
Your notes. (Review of gradient Descent)

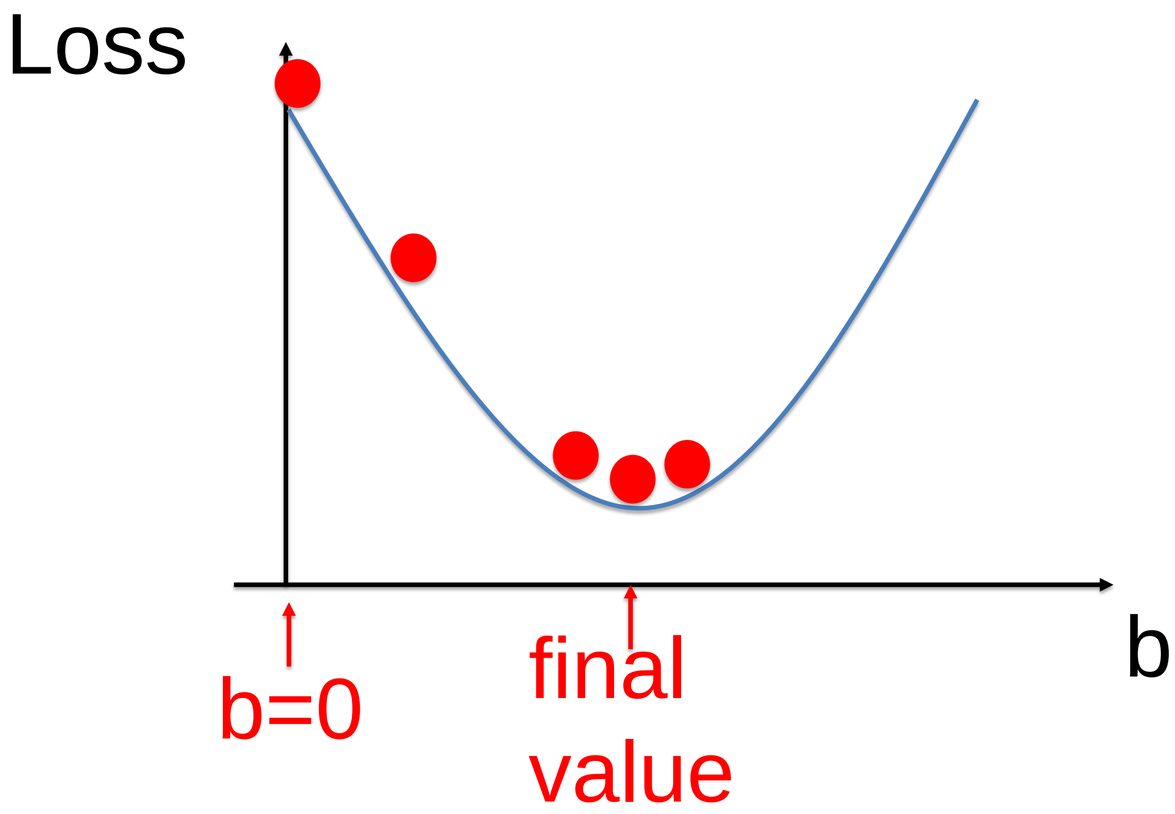
The set of parameters (also called parameter vector) is generically denoted by θ.

The loss function (error function) is denoted by capital L.

.

Detour Machine Learning ReCap: Online, Batch, Expectation





Gradient descent in ‘batch mode’

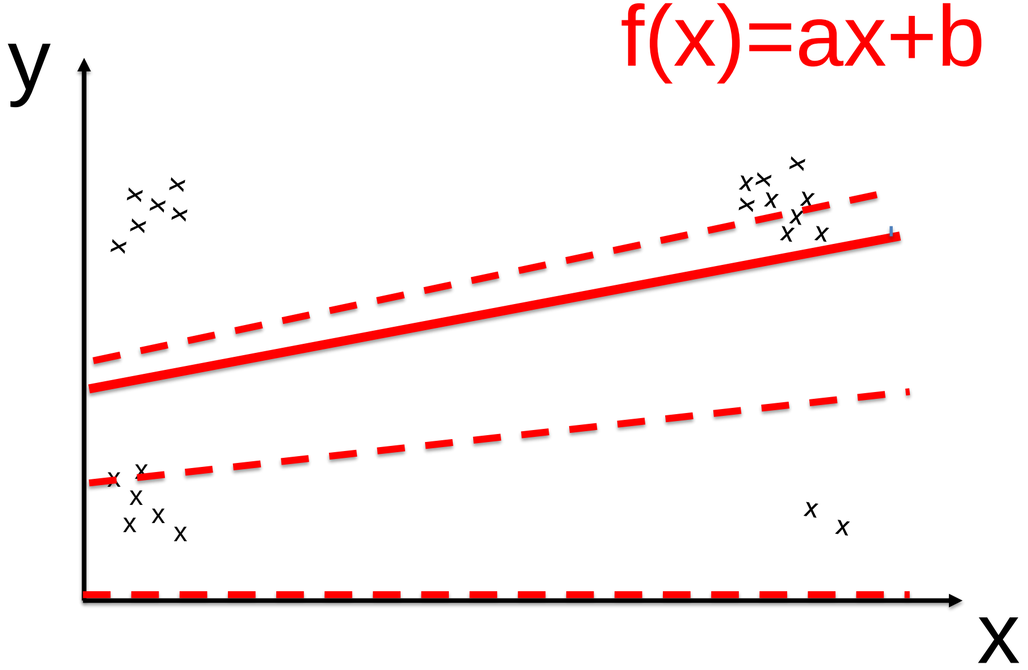
Your notes. (Review of gradient Descent)

The specific parameters of the linear function are a and b. For the drawing of the loss function, only one of the two parameters is plotted.

The aim is to fit a set of data points by the linear function.
Dashed red lines show intermediate update steps.

Note: For the learning rate I will often use the symbol α instead of η

Detour Machine Learning ReCap: Online, Batch, Expectation



$$f(x|\theta) = f(x|a,b) = ax + b$$

Parameters $\theta = (a, b)$

$$\Delta\theta = -\alpha \frac{d}{d\theta} L(\theta)$$

$$\Delta b = -\alpha \frac{\partial}{\partial b} L(a,b)$$

algo (iterative update)

iterate to convergence criteria

$$b^{old} \leftarrow b$$

$$b = b^{old} + \Delta b$$

analogously for a

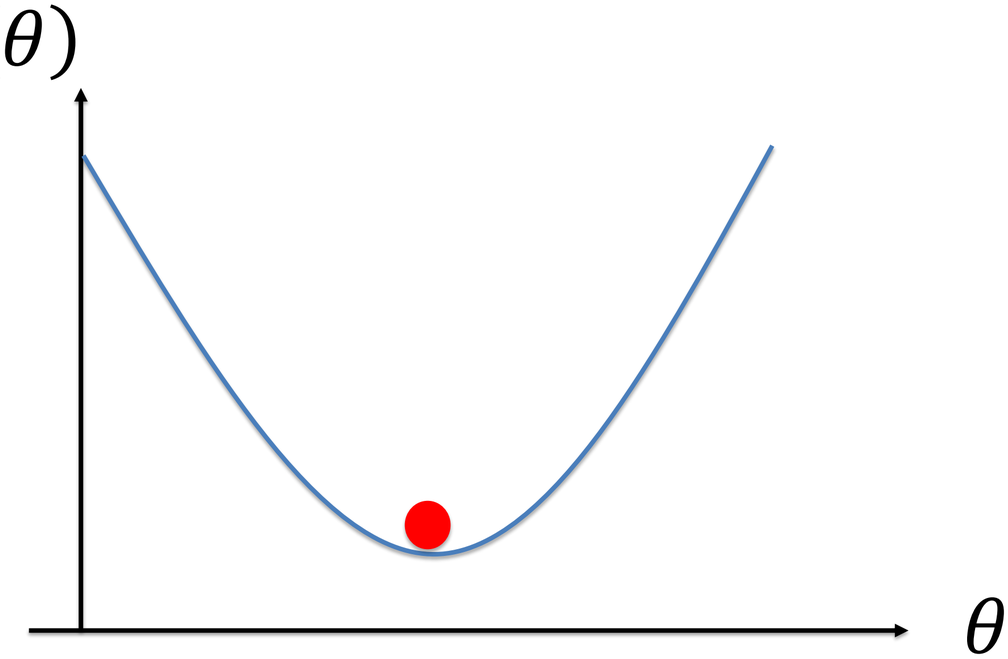
Your notes. (Review of gradient Descent)

The set of parameters (also called parameter vector) is generically denoted by θ.
And then the parameters are specified to be a and b.

Changes of parameters are calculated by gradient descent on the Loss function.

Detour Machine Learning ReCap: Online, Batch, Expectation

CONCLUSION 1, from rule (1):
if $\Delta\theta$=0 then
- **parameter θ no longer changes**
- **(local) minimum at θ^{old}**



Gradient is always evaluated at θ^{old}

$$\Delta\theta = -\alpha \frac{d}{d\theta} L(\theta)|_{\theta^{old}} \qquad (1)$$

algo (iterative update)

iterate to convergence criteria

$$\theta^{old} \leftarrow \theta$$

$$\theta = \theta^{old} + \Delta\theta$$

Your notes. (Review of gradient Descent)

The update rule tells us immediately that the update vanishes at a parameter value that has zero gradient. Only minima can be generically approached by gradient descent (not the maxima).

Detour Machine Learning ReCap: Online, Batch, Expectation

Loss function

$$L(\theta) = \mathbf{E}[l(\mathrm{f}(\mathrm{x}|\theta), \mathrm{y})]$$

$$L(\theta) = \frac{1}{N}\sum_{k}^{N}[l(\mathrm{f}(x_k|\theta), y_k)]$$

loss per data point

Gradient descent (batch)

$$\Delta\theta = -\alpha \frac{d}{d\theta} L(\theta)$$

Example: l =L2 loss, linear model

$$\mathrm{f}(x_k|\theta) = \mathrm{f}(x_k|a, b) = ax_k + b$$

$$L(\theta) = \frac{1}{N}\sum_{k}^{N}(ax_k + b - x_k)^2$$

$$\Delta\theta = -\alpha \, \boldsymbol{E}[\frac{d}{d\theta} \, l(f(x|\theta), y)]$$

$$\Delta\theta = -\alpha \frac{1}{N}\sum_{k}^{N}\frac{d}{d\theta}[l(\mathrm{f}(x_k|\theta), y_k)]$$

Your notes. (Review of gradient Descent)

The loss function is the expectation across all possible pairs (x,y) with the appropriate statistical weight. The loss per data point is denote by a small character l .

Often a large batch of N data points is taken instead. These N data points must be representative for the statistical distribution p(x,y).

The example shows the linear function that we considered earlier.

Detour Machine Learning ReCap: Online, Batch, Expectation

Loss function

$$L(\theta) = \mathbf{E}[l(\mathrm{f}(\mathrm{x}|\theta), \mathrm{y})]$$

$$L(\theta) = \int dxdy\, p(x,y)[l(\mathrm{f}(\mathrm{x}|\theta), \mathrm{y})]$$

$$L(\theta) = \frac{1}{N}\sum_{k}^{N}[l(\mathrm{f}(x_k|\theta), y_k)]$$

loss per data point

Gradient descent (batch)

$$\Delta\theta = -\alpha \frac{d}{d\theta} L(\theta)$$

$$\Delta\theta = -\alpha \int \boldsymbol{dxdy}\, \boldsymbol{p}(\boldsymbol{x},\boldsymbol{y})[\frac{d}{d\theta}\, l(f(x|\theta), y]$$

$$\Delta\theta = -\alpha\, \boldsymbol{E}[\frac{d}{d\theta}[l(f(x|\theta), y)]]$$

$$\Delta\theta = -\alpha \frac{1}{N}\sum_{k}^{N}\frac{d}{d\theta}[l(\mathrm{f}(x_k|\theta), y_k)]$$

Your notes. (Review of gradient Descent)

On the left:
The loss function is the expectation across all possible pairs (x,y) with the appropriate statistical weight. The gradient operation is linear and can be exchanged with the expectation (which is also a linear operation).

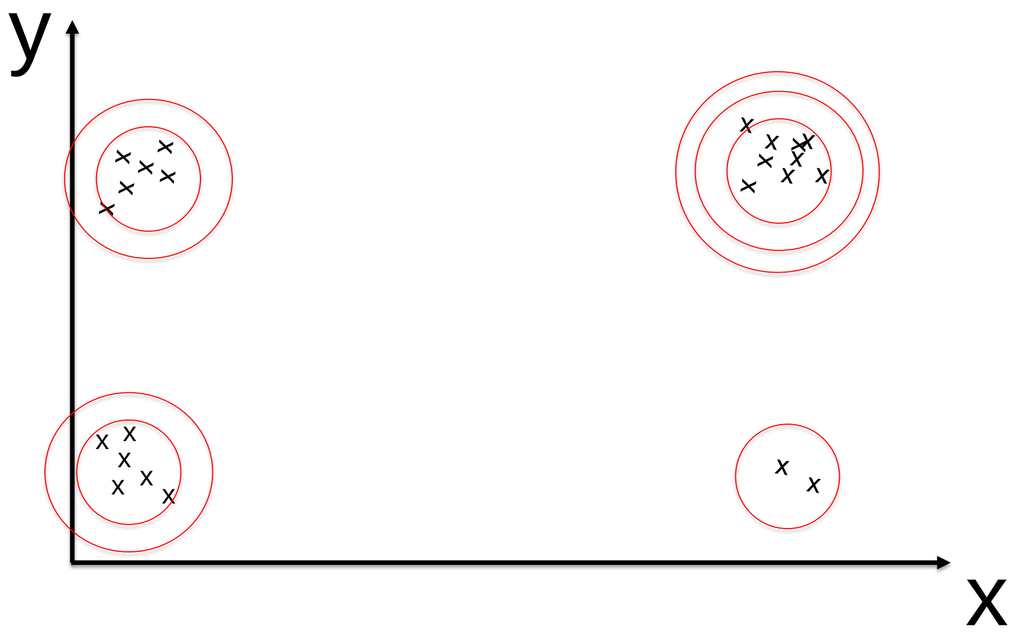
On the right:
The same calculation with a large batch of N data points.
The average of N in the gradient is analogous to the expectation (red boxes).

Conclusion: Expectation = Batch size N to infinity

$$\Delta\theta = -\alpha\, \boldsymbol{E}[\frac{d}{d\theta}\, l(f(x|\theta), y]$$

$$\Delta\theta = -\alpha \int \boldsymbol{dxdy}\, \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y})[\frac{d}{d\theta}\, l(f(x|\theta), y]$$

‘statistical formulation’ with ‘expectations’



choose N data points using the appropriate statistical weight

$$\Delta\theta = -\alpha\, \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k}}^{\boldsymbol{N}} \frac{d}{d\theta} [l(\mathrm{f}(x_k|\theta), y_k)]$$

$$\boldsymbol{E}[\dots] = \lim_{N\to\infty} \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k}}^{\boldsymbol{N}} [\dots]$$

Your notes. (Review of gradient Descent)

We said that the average of N in the gradient is ‘analogous’ to the expectation.
For the limit N to infinity batch and expectation are again identical.

The idea of the density p(x,y) is shown for the same example as before.

Conclusion: Expectation = Batch size N to infinity

$$\Delta\theta = -\alpha\, \boldsymbol{E}[\frac{d}{d\theta}\, l(f(x|\theta), y]$$

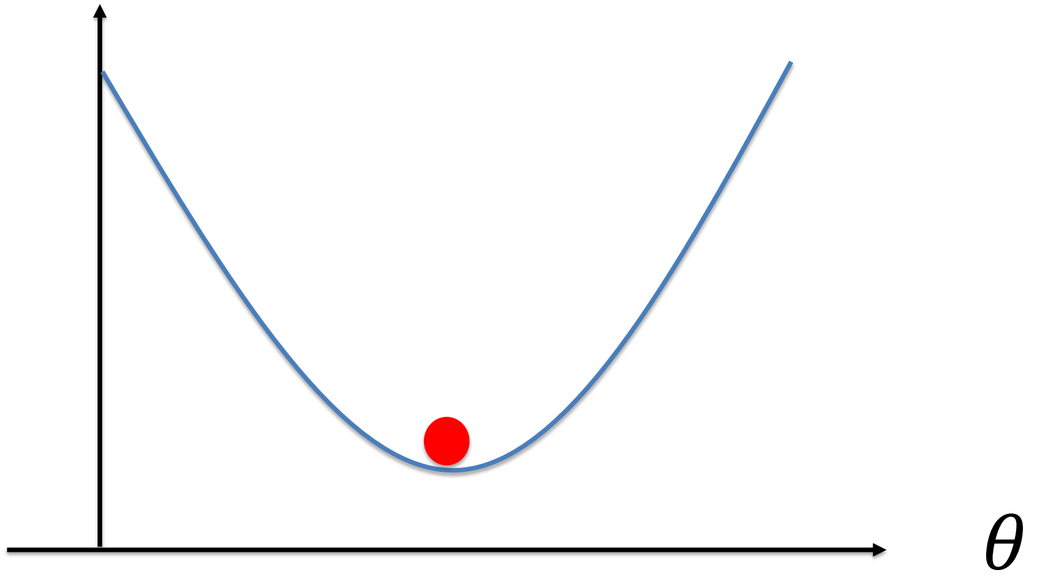
$$\boldsymbol{E}[\dots] = \lim_{N\to\infty} \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k}}^{\boldsymbol{N}} [\dots]$$

$$\Delta\theta = -\alpha\, \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k}}^{\boldsymbol{N}} \frac{d}{d\theta} [l(\mathrm{f}(x_k|\theta), y_k)]$$

CONCLUSION 1 from rule (1):

if $\Delta\theta$=0 (with N to infinity) then

- **$\boldsymbol{\theta}$ doesn't change**
- **(local) minimum at $\boldsymbol{\theta^{old}}$**
- **$\boldsymbol{\theta^{old}}$=$\boldsymbol{\theta^{optim}}$ in 'statistical' sense**



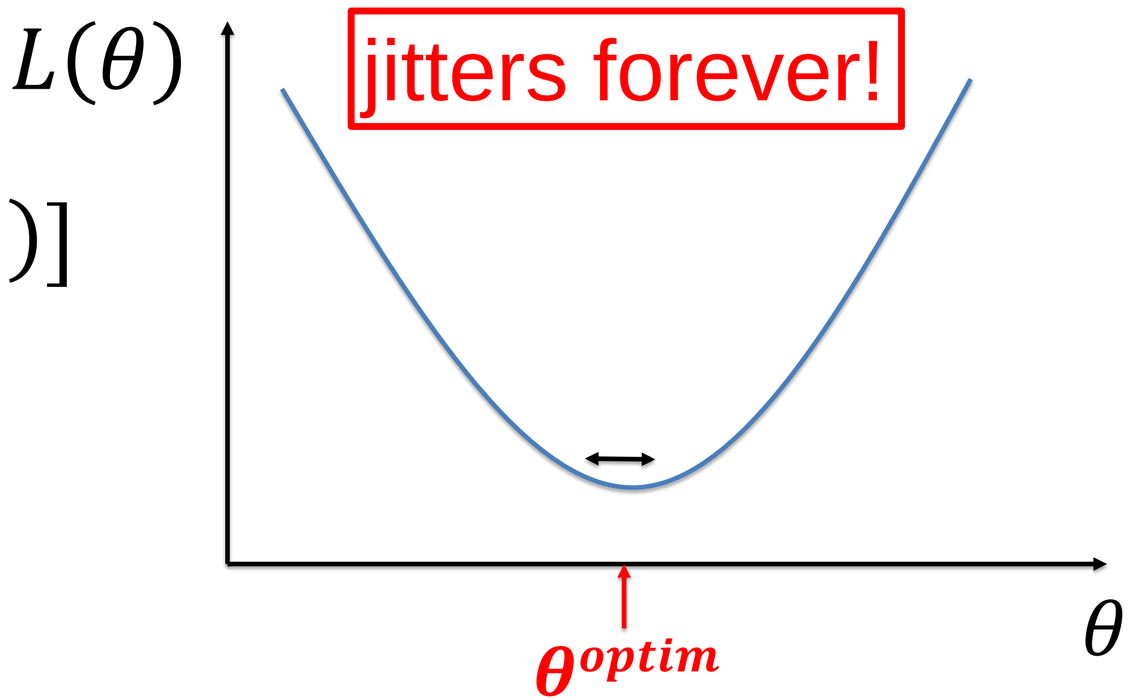
Your notes. (Review of gradient Descent)

A repetition of what we have seen before:
If the update step in the batch rule (N to infinity) vanishes, then we know that we are at a minimum of the loss function.

And this is equivalent to saying:
If the update step of the true loss function with the expectation sign vanishes, then we know that we are at a minimum of the loss function.

Detour Machine Learning ReCap: Batch versus 'Online'

$$\Delta\theta = -\alpha \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k}}^{\boldsymbol{N}} \frac{d}{d\theta} [l(\mathrm{f}(x_k|\theta), y_k)]$$



Online:

$$\Delta\theta = -\alpha \frac{d}{d\theta} [l(\mathrm{f}(x_k|\theta), y_k)]_{\theta=\theta^{old}}$$

Update after each data point

algo (iterative update)

iterate to convergence criteria

$$\theta^{old} \leftarrow \theta$$

$$\theta = \theta^{old} + \Delta\theta$$

Your notes. (Review of gradient Descent)

An online rule means: drop the statistical averaging.

Here it means: drop the sum over data points.

As a result the parameter vector θ
can change after each data point!

And this is true even if we are already at the exact minimum of the true loss. The next data point might for example be an outlier and the parameter vector changes again. Therefore the gradient descent solution always jitters.

The size of the jitter depends on the learning rate (here called alpha).

Detour Machine Learning ReCap: Batch, 'Online', Expectation

Online:

$$\Delta\theta = -\alpha \frac{d}{d\theta}[l(\mathrm{f}(x_k|\theta), y_k)]_{\theta=\theta^{old}}$$

Update after each data point

Conclusion:

- Online update has jitter

BUT

- Expected update has no jitter

Expected Online Update ($\theta = \theta^{old}$ frozen):

$$E[\Delta\theta] = -\alpha E[\frac{d}{d\theta}[l(\mathrm{f}(x_k|\theta), y_k)]_{\theta=\theta^{old}}]$$

Conclusion:

Expected update of the online rule is identical to batch update with infinite data

Your notes. (Review of gradient Descent)

We can ask:
What would be the EXPECTATION of the update step.

Suppose we momentarily have the parameter $\theta = \theta^{old}$.
Then we ask what is the EXPECTED change at this location.

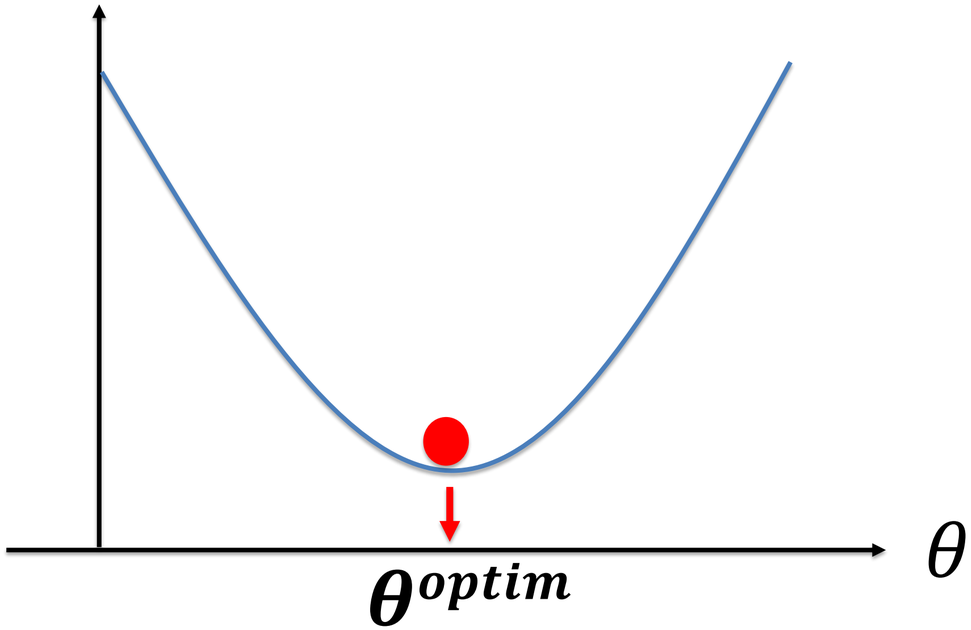
Comparison with the batch rule shows that the expected update of the online rule is identical to batch update with infinite data evaluated at $\theta = \theta^{old}$.

THIS IMPLIES:
If by chance $\boldsymbol{\theta = \theta^{old}}$ is the exact minimum, the expected update is zero; but the ACTUAL update can be nonzero!

This is also summarized in the next slide and the quiz.

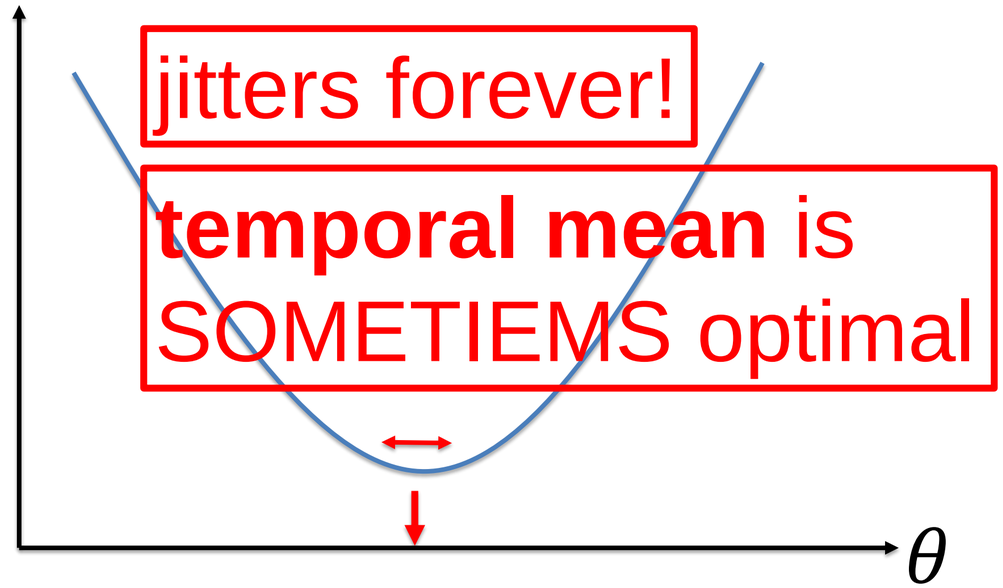
Batch rule with N to infinity

CONCLUSION 1 from rule (1):
if $\Delta\theta$=0 (with N to infinity) **then**
- $\boldsymbol{\theta}$ **doesn't change**
- **(local) minimum at** $\boldsymbol{\theta^{old}}$
- $\boldsymbol{\theta^{old}}$**=**$\boldsymbol{\theta^{optim}}$



Online Rule

CONCLUSION 2:
- $\boldsymbol{\theta}$ **jitters forever. BUT:**
- **if by chance** $\boldsymbol{\theta^{old}}$ **such that** $E(\Delta\theta)$=0 **then**
(local) minimum at $\boldsymbol{\theta^{old}} = \boldsymbol{\theta^{optim}}$



Previous slide/next slide. Summary slides:

If the expected update is zero [i.e., $\boldsymbol{E}(\Delta\theta)$=0] for a given set of parameter $\theta = \boldsymbol{\theta^{old}}$, then θ is a locally optimal parameter, **even for the online rule**:

$$\theta = \boldsymbol{\theta^{old}} = \boldsymbol{\theta^{optim}}$$

There is no statement how we would find this parameter $\theta = \boldsymbol{\theta^{optim}}$

A completely different statement concerns the **mean of the jittering parameter** $\boldsymbol{\theta}$.

If the **update steps are symmetric**, then the mean of the parameter θ is the optimal one: $\langle\theta\rangle = \boldsymbol{\theta^{optim}}$

However, if the update steps are asymmetric, then the mean $\langle\theta\rangle$ of the parameter θ is shifted compared to the optimal one (next slide).

For gradient descent on a loss function we recognize the asymmetry in loss curve.

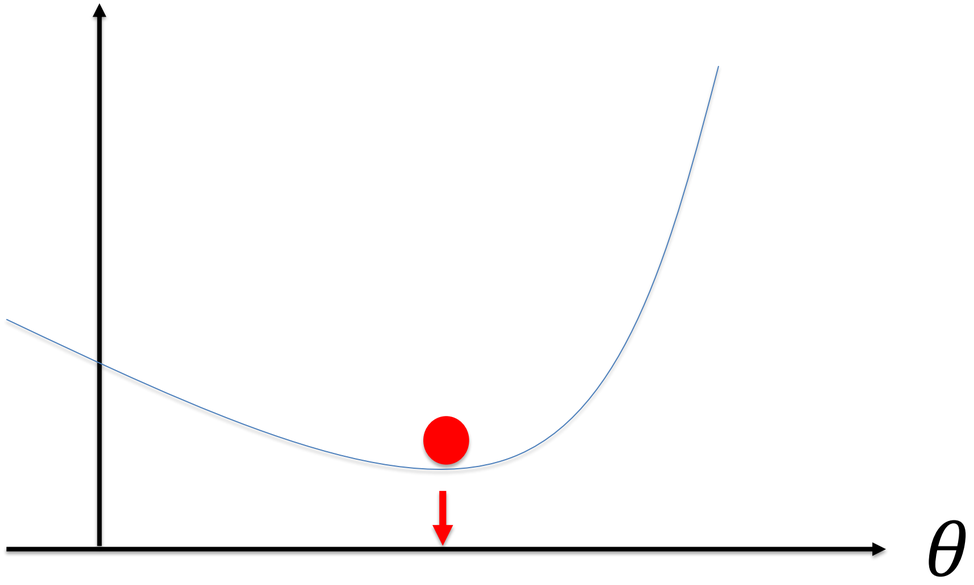
However (even in cases where we do not have a loss function) what really counts is whether the update steps are symmetric or not:

Suppose the current parameter is $\theta = \boldsymbol{\theta^{optim}} + \epsilon$ where is small, i.e. close to the optimum

Symmetry is guaranteed if update steps are linear $\Delta\theta = \alpha\epsilon$ with small constant α.

Batch rule with N to infinity

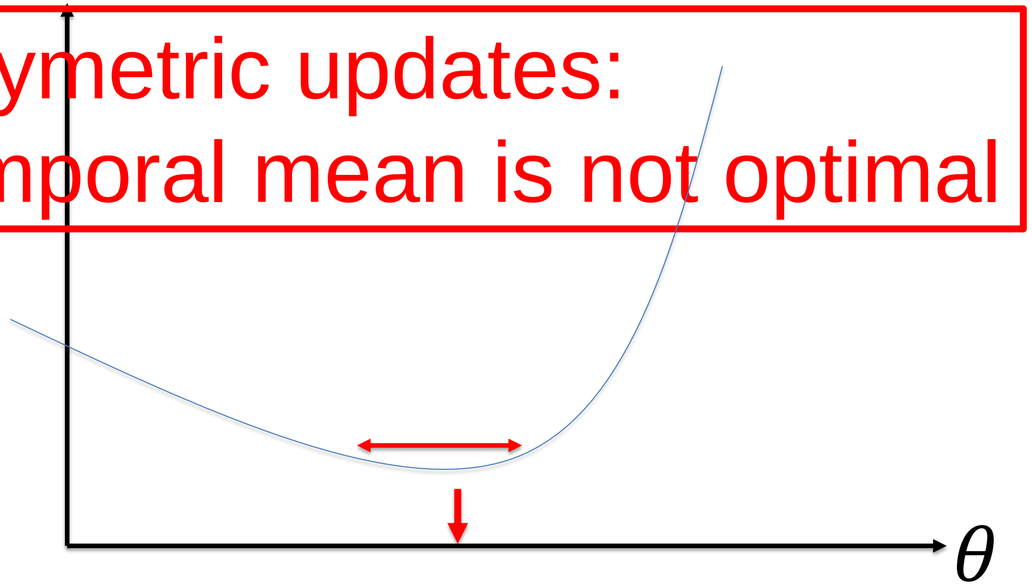
CONCLUSION 1 from rule (1):
if $\Delta\theta$=0 (with N to infinity) **then**
- $\boldsymbol{\theta}$ **doesn't change**
- **(local) minimum at** $\boldsymbol{\theta^{old}}$
- $\boldsymbol{\theta^{old}}$**=**$\boldsymbol{\theta^{optim}}$



Online Rule

CONCLUSION 2:
- $\boldsymbol{\theta}$ **jitters forever. BUT:**
- if by chance $\boldsymbol{\theta^{old}}$ **such that**
$E(\Delta\theta)$=0 **then**
(local) minimum at $\boldsymbol{\theta^{old} = \theta^{optim}}$

Asymetric updates:
temporal mean is not optimal



Quiz: Expectation, Batch, Online (Recap of ML)

[ ] With a **batch rule** and small learning rate, I sometimes reach a local minimum without remaining parameter jitter.

[ ] With a **batch rule at a local minimum** I never have any remaining parameter jitter

[ ] With an **online rule at a local minimum** I never have any remaining parameter jitter

[ ] With an **online rule** at a local minimum the **expectation of the online update step vanishes.**

[ ] The **expectation of the online update step** is equivalent to a very large **batch** (N to infinity)

[ ] With an online rule jittering round the minimum, the temporal mean is guaranteed to be at the location of the minimum

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 3c: Apply Expectation of ONLINE to Q-values

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step horizon (bandit problems)
- **Expectation, batch, and online rules**

End of Detour:
Apply to Q-values in the Bandit problem.

ML: parameters are called θ

Function fitting: parameters are a and b

Bandit problem: parameters are Q(s,a)

Recall: Update rule in Expectation (weak version)

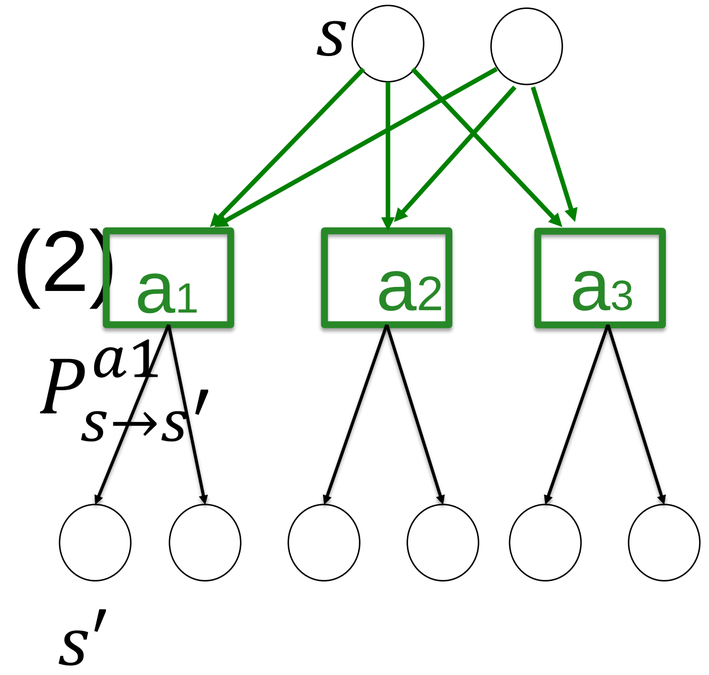
After taking action a in state s, we update with

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)] \qquad (1)$$

(i) If the expectation (over s', r) of the update rule (1) **given (s,a)** vanishes**,** then $\hat{Q}(s,a)$ has an expectation value,

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

(ii) If the learning rate η decreases, fluctuations around the **empirical mean** $\langle \hat{\boldsymbol{Q}}(\boldsymbol{s},\boldsymbol{a})\rangle$ decrease and **the empirical mean approaches** $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$



Previous slide.

When evaluating the **expectation value given (s,a),** the learning rate drops out since we set the left-hand-side to zero. The exact value of η is not relevant, as discussed in the theorem. Part (i) of the theorem states that the expectation value of $\hat{Q}(s,a)$ is the correct Q-value. For a quick proof of part (i) see the video. On the blackboard a stronger statement was shown.

Convergence in expectation is equivalent to imagining that you start millions of trials with the same value $\hat{Q}(s,a)$ without any intermediate update. So in that sense it is like a super-big 'batch' of examples.

In practice, we do not have expectations but online updates with fluctuations. It is important that the learning rate η is small at the end of learning so as to limit the amount of fluctuations. Part (ii) states that **online mean** for small learning rate also goes to the correct Q-value.

Indeed, since the equations are linear (for the bandit problem = 1-step horizon), the calculation of part (i) apply analogously to the long-term empirical temporal average (denoted by angular brackets) $\langle \Delta \hat{Q}(s,a) \rangle = \eta \langle [r_t - \hat{Q}(s,a)] \rangle$

This equivalence based on linearity is not true for the multi-step horizon that we discuss later in this lecture.

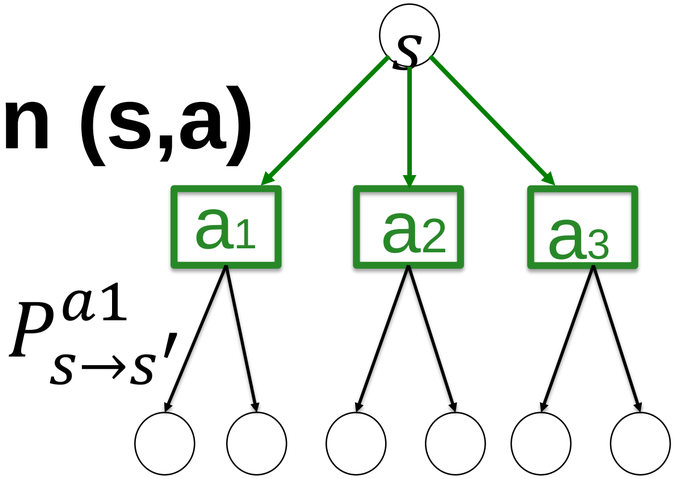
Recall: Proof for strong version (blackboard last week)

After taking action a in state s, we update with

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)] \qquad (1)$$

(i) If the expectation of the update rule (1) **given (s,a)** vanishes, then $\hat{Q}(s,a)$ has an expectation

$$E\left[\hat{Q}(s,a)\right] = \hat{Q}(s,a) = \sum_{s'} P^{a}_{s\to s'}\, R^{a}_{s\to s'} = Q(s,a) \qquad (2)$$



'I can even drop the expectation sign in the results'

If updates $E(\Delta\hat{Q}(s,a)|s,a)=0$

then $\hat{Q}(s,a)$ is the correctt solution.

' $\hat{Q}(s,a)$ takes role of $\boldsymbol{\theta^{old}}$ '

Your notes.

When I loosely say: ‘converged in expectation’ then I simply mean that the **expectation of the update rule vanishes.**

So the interpretation is:
if by chance we found a parameter $\hat{Q}(s,a)$
that fulfills $E(\Delta\hat{Q}(s,a)|s,a)=0$

then we know that $\hat{Q}(s,a)$ is correct.

The theorem does not say how we would find it: we do not know whether it is stable under the online update algorithm.

Part (i) of Theorem (strong version)

Expectation of online rule=0 → $E(\Delta\hat{Q}(s,a)|\text{s,a})=0$

expectation of all possible futures with correct statistical weight

we always start in (s,a) while the system is frozen;
→ $\hat{Q}(s,a) = Q^{old}$ frozen

Update using **expectation over all possibilities that may occur in the next time step**.

= Perspective similar to a batch mode:

update only **after** (infinitely) many samples = trials that all start in (s,a) with the same frozen value Q^{old}

$$E(\Delta\hat{Q}(s,a)|\text{s,a}) = \eta\left[E(r_t) - Q^{old}\right] = \lim_{N} \frac{1}{N}\sum_{k=1}^{N} \eta\left[r_t - Q^{old}\right]$$

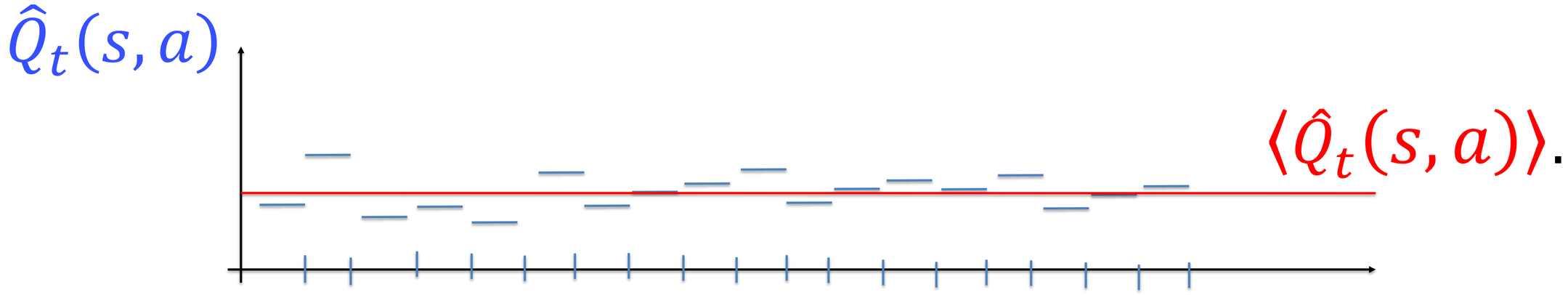
Previous slide:
$\hat{Q}(s,a)$ denotes the current estimate of the Q-value. Claim: If Q no longer changes (in expectation) then it must be the correct Q-value.

There are different views on how to interpret the ‘expectation;:
- Formally from a mathematical point of view: average over all possible outcomes of the next time step given (s,a).
- In a simulation this would correspond to the following sampling procedure: You freeze the value of $\hat{Q}(s,a)$ and run MANY times (N to infinity) a test with the state-action pair (s,a) as a starting condition. Then you evaluate the resulting ‘batch update’ averaged across all these examples. If the batch update with Millions of Examples is zero, that implies that you have converged to the correct value.

Blackboard3

Part (ii) of Theorem:

We work with the **online update** $\Delta\hat{Q}(s,a)$. With finite learning rate, the value of $\hat{Q}_t(s,a)$ fluctuates around a mean $\langle\hat{Q}_t(s,a)\rangle$.



Claim: Under the hypothesis $\langle\Delta\hat{Q}(s,a)=0\rangle$, the mean $\langle\hat{Q}_t(s,a)\rangle$ is equal to the 'correct' Q-value.

Your notes. (Proof in the Blackboard notes)

Part (ii) of Theorem:

Claim: Under the hypothesis $\langle \Delta \hat{Q}(s,a) = 0 \rangle$, the temporal mean $\langle \hat{Q}_t(s,a) \rangle$ with online updating is equal to the 'correct' Q-value $Q(s,a)$.

Proof:

$$\Delta \hat{Q}(s,a) = \eta[r_t - \hat{Q}_t(s,a)] \qquad (1)$$

$$\langle \Delta \hat{Q}(s,a) \rangle = \eta \langle r_t - \hat{Q}_t(s,a) \rangle = 0$$

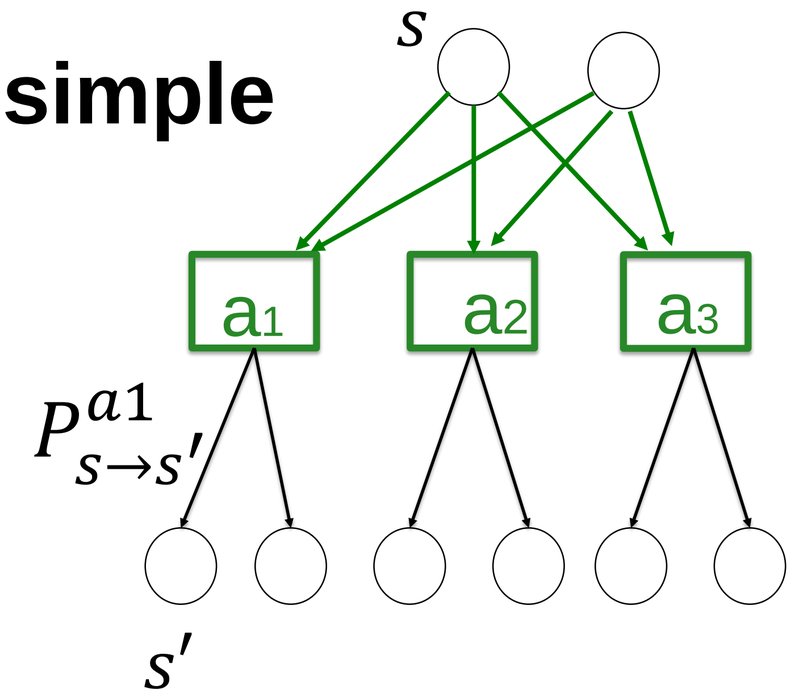
$$0 = \langle r_t \rangle - \langle \hat{Q}_t(s,a) \rangle$$

$\langle r_t \rangle = Q(s,a)$ by definition of Q-values

Your notes. (Proof in the Blackboard notes)

One-step horizon: summary

Q-value = expected reward for state-action pair



If Q-value is known, choice of action is simple

→ take action with highest Q-value

If Q-value not known:

→ estimate $\hat{Q}$ by trial and error

→ update with rule

$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)\,] \qquad \textbf{(1)}$$

→Let learning rate η decrease over time

- **Iterative algorithm (1) always fluctuates for finite η**
- **The expectation of the update step is 'good'**

Previous slide.
Let us distinguish the ESTIMATE $\hat{Q}(s,a)$ from the real Q-value $Q(s,a)$

The update rule can be interpreted as follows:
if the actual reward is larger than (my estimate of) the expected reward, then I should increase (a little bit) my expectations.

The learning rate η :

In exercise 1, we found a rather specific scheme for how to reduce the learning rate over time. But many other schemes also work in practice. For example you keep η constant for a block of time, and then you decrease it for the next block.

Note: I will often use the symbol α instead of η

Both symbols indicate what is called the ‘learning rate’ in Deep Learning.

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 4: Exploration vs. Exploitation

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step horizon (bandit problems)
- **Exploration vs. Exploitation**

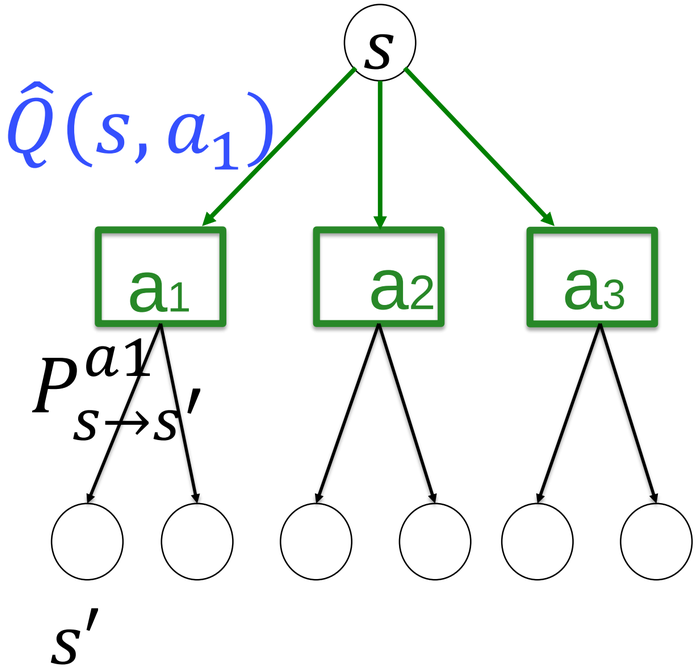
Previous slide.

To estimate the Q-values you have to play all the different actions several times. However, if you know the Q-values you should only play the best action.

Exploration – Exploitation dilemma

Ideal: take action with maximal $Q(s,a)$

Problem: correct Q values not known
(since reward probabilities and branching probabilities unknown)



Exploration versus exploitation

Explore so as to estimate reward probabilities

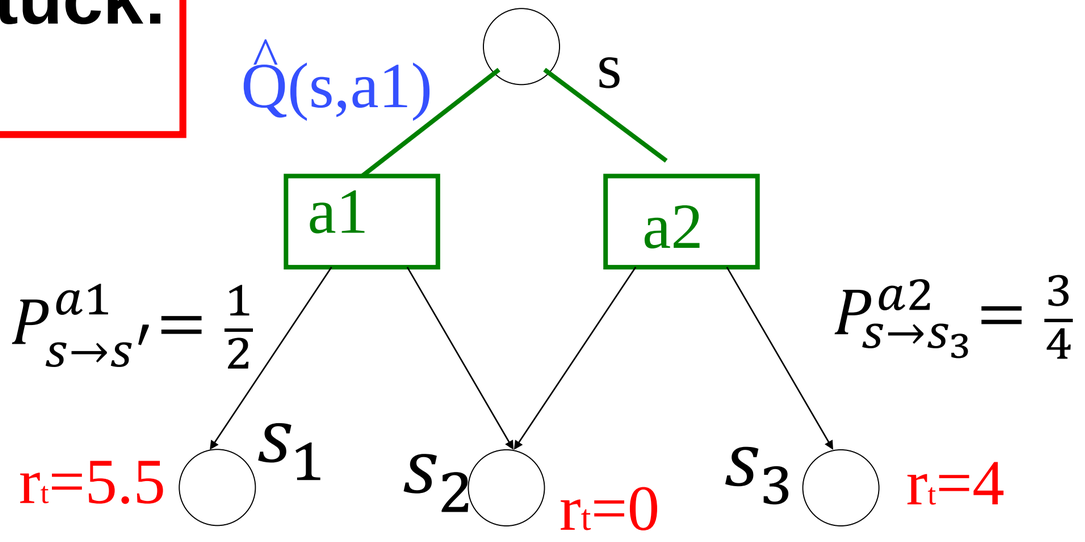
Take action which looks optimal, so as to maximize reward

Previous slide.

Since Q-values are not known, you are always in the situation of an exploration-exploitation dilemma.

Note: All estimates of Q will be empirical estimates. Here in this and the next slide I still write Q-hat for the empirical average. However, later, I simplify the notation and write for the empirical estimate Q(s,a) without the hat whenever the meaning is implicitly clear.

greedy makes you stuck: Example



s=state

a2 action

r_t=1 actual reward

s'=new state

Assume that you initialize all Q values with zero; set η =0.2 (constant)

update $\Delta\hat{Q}(s,a) = \eta[r_t - \hat{Q}(s,a)]$

Trial 1: you choose action a1, you get r_t=5.5

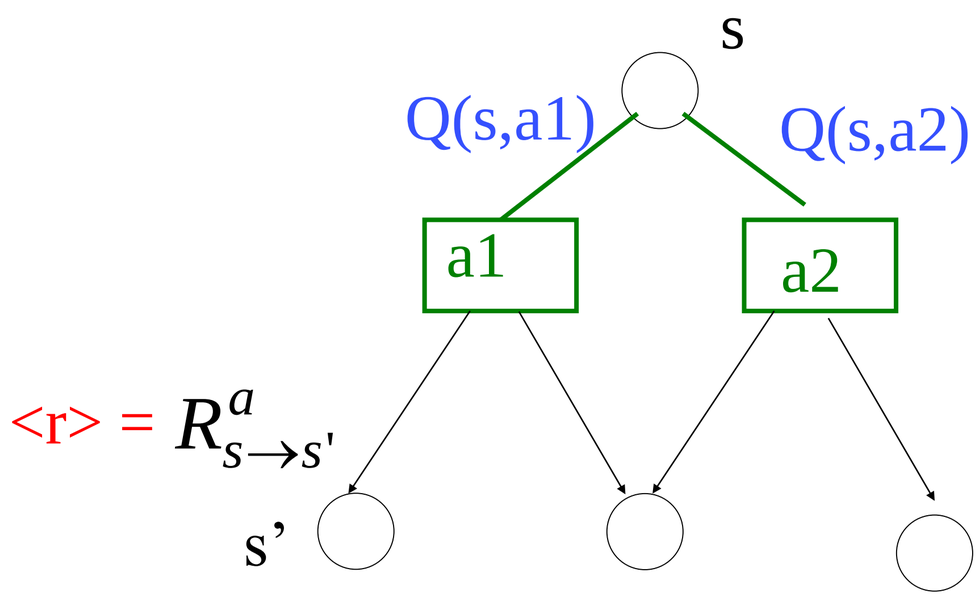
Trial 2: you choose action a2, you get r_t=4.0

Trial 3 – 4: continue **'greedy': → you continue with action 1**

BUT: the expected reward Q(s,a)= $\sum_{s'} P^{a}_{s\to s'} R^{a}_{s\to s'}$ is larger for action 2.

Given the outcomes of the first two trials, action a1 looks better.
You can check that whatever the outcome in trial 3 (even for reward=0!), the estimated Q-value of action a1 is still higher than that of action a2!

Exploration and Exploitation



Problem: correct Q values not known

greedy strategy:
- take action a* which looks best

$$Q(s,a^*) \geq Q(s,a_j) \quad \text{for all } j$$

ATTENTION:
with ‘greedy’ you may get stuck with a sub-optimal strategy (see Exercise!)

Previous slide.

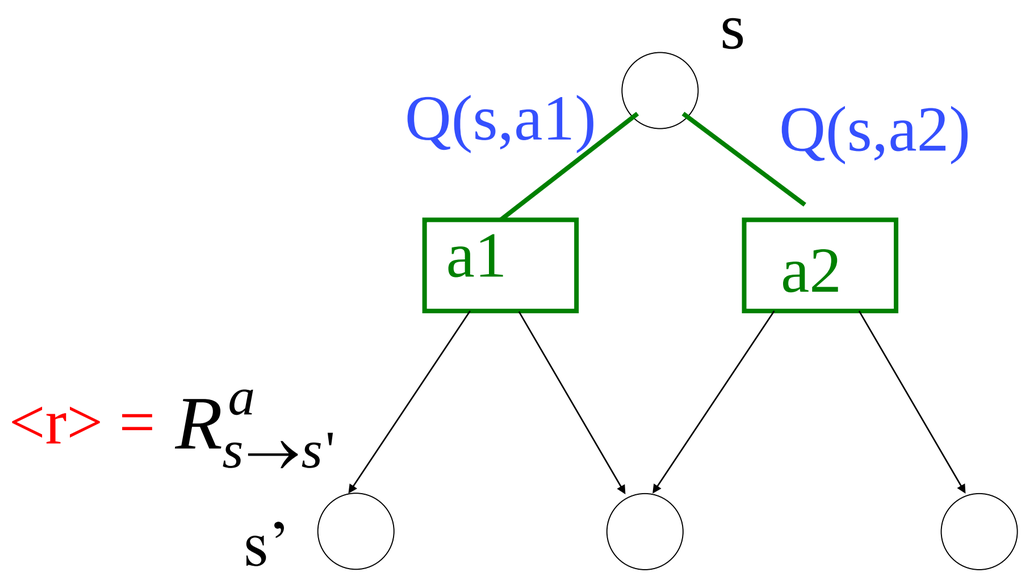
If you know the correct Q-values, the best choice would be to choose the action with maximal Q-value (called ‘greedy’ action). But since you don’t know the Q-values it is risky to choose the greedy action because you may get stuck with a suboptimal choice.

In (almost all) applications of reinforcement learning we work with estimated Q-values.

Previously we used a hat to distinguish the ESTIMATED $\hat{Q}(s,a)$ from the real Q-value $Q(s,a)$. However, in the following I will write the estimated Q-values without the hat. Nearly always Q means estimated Q.

Exploration and Exploitation: practical approach

hats have been dropped!



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

Problem: correct Q values not known

greedy strategy:

- take action a* which looks best

$$Q(s,a^*) \geq Q(s,a_j) \text{ for all } j$$

$\mathcal{E}$greedy strategy:

- take action a* which looks best with prob $P = 1-\varepsilon$

Softmax strategy: take action a' with prob

$$P(a') = \frac{\exp[\beta Q(a')]}{\sum_a \exp[\beta Q(a)]}$$

Optimistic greedy:

initialize with Q values that are too big

Previous slide.

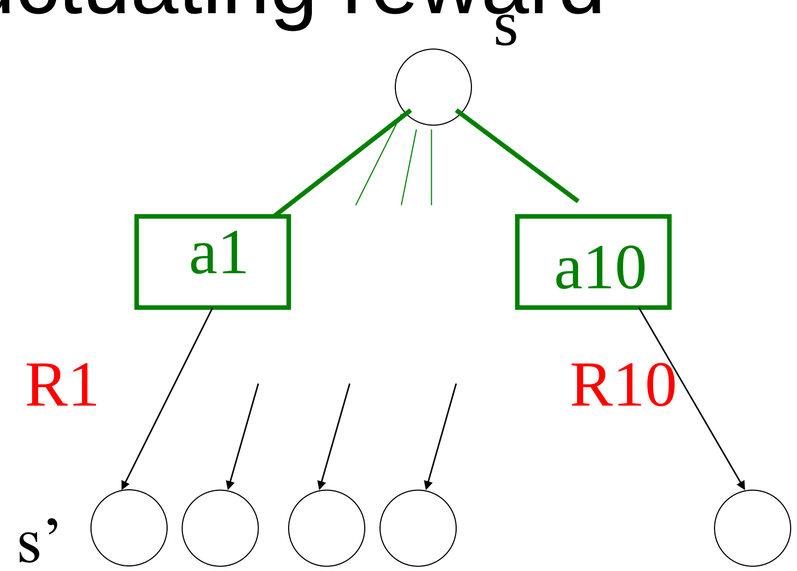
Softer versions of greedy allow you to choose occasionally an action which looks suboptimal, but which allows you to further explore the Q-values of other options.

Epsilon-greedy and softmax are examples following this idea.
Note that ‘softmax’ is a function that one also encounters in multiclass tasks with 1-hot coding (see course of ‘machine learning’; also later lecture on deep learning)

A radically different approach is optimistic greedy. If you initialize all Q-values at the same value, but clearly too high (compared to maximal reward that you can get in the scheme), then the Q-value of action a1 decreases initially each time you play a1, which in turn favors other actions that you have not yet played.

Exploration and Exploitation: practical approach

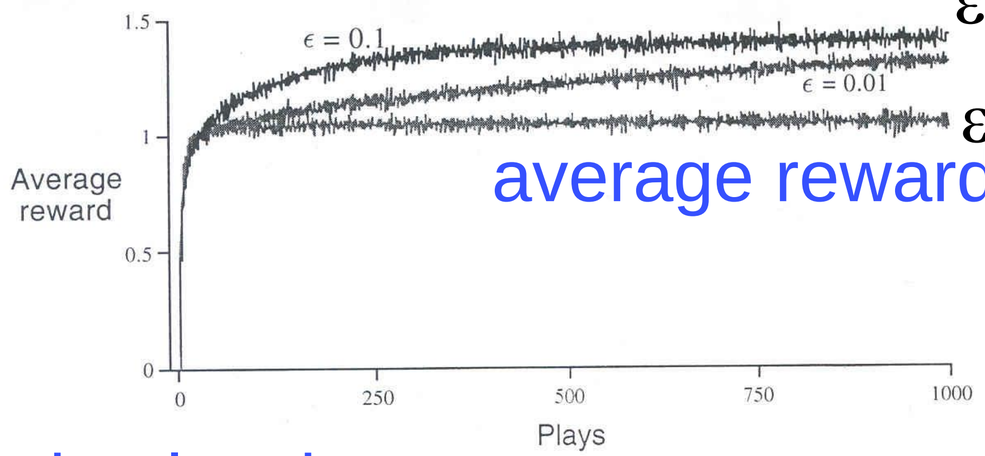
Example: 10-armed bandit with fluctuating reward



in each action, actual rewards fluctuate around a mean

Rk= $R_{s \to s'}^{a_k}$

Epsilon-greedy: simulation



Optimal action

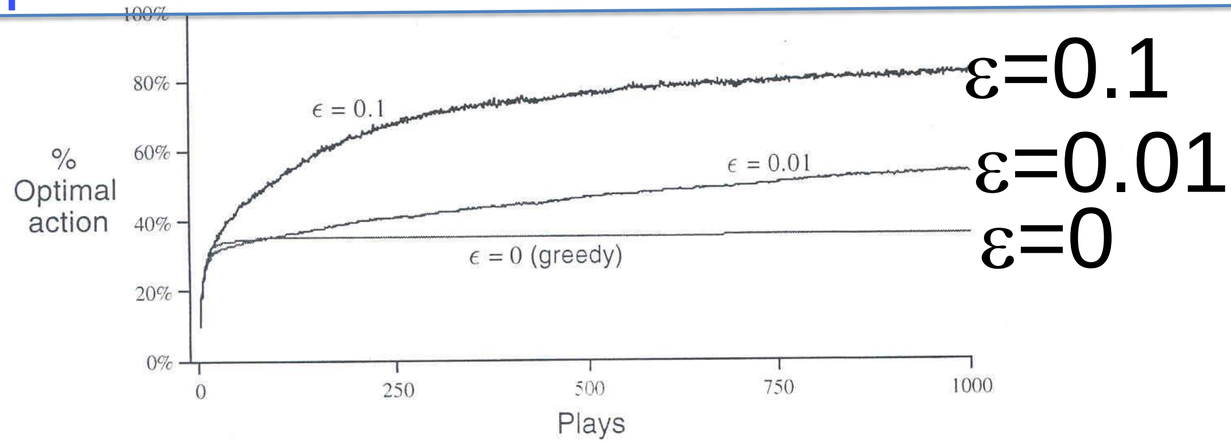


Figure 2.1 Average performance of ϵ-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 tasks. All methods used sample averages as their action-value estimates.

book: Sutton and Barto

Previous slide.

Computer simulation of a situation where actual rewards r fluctuate around the mean reward R. There are 10 different actions a1, …, a10 each with a different mean reward R1, …, R10.

There exist two different ways to evaluate the performance.
Top: what is the average reward that you get by playing epsilon-greedy?
Bottom: what is the fraction of times that you play the optimal action, by playing epsilon-greedy.

Three different values of epsilon are used.

Exploration and Exploitation: practical approach

Epsilon-greedy, combined with iterative update of Q-values

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Repeat forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

Sutton and Barto call R what I call r_t

learning rate

Sutton and Barto, ch. 2

Previous slide.

This is the style of pseudo-code that we will see a lot over the next few weeks. It is taken from the book of Sutton and Barto (MIT Press, 2018); Sutton and Barto made a pdf online available.

Q(a) is the Q-value for action a. Since we have always the same starting state in which we have to make our choice of action, we can suppress the index of the state s. Q(a) = Q(s_{start},a).

N(a) is a counter of how many times the agent has taken action a.

In this specific example the learning rate eta is the inverse of the count N(a) (see earlier exercise); but in the more general setting we would remove the counter and just use some heuristic reduction scheme for eta.

Note that in class we define (1-epsilon) as the probability of taking the ‘best’ action corresponding to argmax Q and epsilon is then distributed over the OTHER actions. Sutton and Barto distribute epsilon over ALL actions, including the ‘best’.
Thus for a total choice of 3 actions, Sutton and Barto have a probability of epsilon/3 for the other actions (and with the definition in class it would be epsilon/2).

Quiz: Exploration – Exploitation dilemma

We use an iterative method and update Q-values with **eta=0.1**

[ ] With a greedy policy the agent uses the best possible action

[ ] Using an epsilon-greedy method with epsilon = 0.1
means that, even after convergence of Q-values,
in about 10 percent of cases a suboptimal action is chosen.

[ ] If the rewards in the system are between 0 and 1 and Q-values are initialized with Q=2, then each action is played at least 5 times before exploitation starts.
(exploitation starts when you no longer choose the wrong action)

Previous slide.

Here we define as in class (1-epsilon) as the probability of taking the ‘best’ action corresponding to argmax Q and epsilon is then distributed over the OTHER actions.

Quiz: Exploration – Exploitation with Softmax policy

Softmax policy: take action a' with prob $P(a') = \frac{\exp[\beta Q(a')]}{\sum_a \exp[\beta Q(a)]}$

[ ] Suppose we have 3 possible actions a_1, a_2, a_3 and use the softmax policy. Is the following claim true?
For $Q(a_1) = 4, Q(a_2) = 1, Q(a_3) = 0$ the preference for action a_1 is more pronounced than for $Q(a_1) = 34, Q(a_2) = 31, Q(a_3) = 30$.

Quiz: Exploration – Exploitation with Softmax policy

All Q values are initialized with the same value Q=0.1

Rewards in the system are r =0.5 for action 1 (always)
and r=1.0 for action 2 (always)

We use an iterative method and update Q-values with **eta=0.1**

1. [ ] if we use softmax with **beta = 10**, then, after 100 steps, action 2 is chosen almost always
2. [ ] if we use softmax with **beta = 0.1**, then, after 100 steps action 2 is taken about twice as often as action 1.

Softmax policy: take action a'
with prob $P(a') = \frac{\exp[\beta Q(a')]}{\sum_a \exp[\beta Q(a)]}$

Your notes (Quiz not given in class).

Use that exp(5) is a big number!

[yes], since beta[Q(a2)-Q(a1)]=5

[no], with beta=0.1, exp(beta*Q)=1+…
→both action chosen with about the same prob.

Softmax policy: take action a’ with prob

$$P(a') = \frac{\exp[\beta Q(a')]}{\sum_{a} \exp[\beta Q(a)]}$$

Exploration and Exploitation: Summary

Now Exercise 1* - next lecture at 14h15

- If we know the Q-values we can exploit our knowledge
- Exploitation = action which is best = argmax Q(a)

- But we never know the Q-values for sure
- We need to estimate the Q-values by playing the game
- Explore possibilities, transitions, outcomes, reward

For complex problems, there is no perfect trade-off between exploration and exploitation

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 5: Bellman Equation

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step Horizon (Bandit Problems)
- Exploration vs. Exploitation
- **Bellman Equation**

Previous slide.
So far our Q-values were limited to situations with a 1-step horizon. Now we will get more general.

Multistep horizon

Policy $\pi(s,a)$

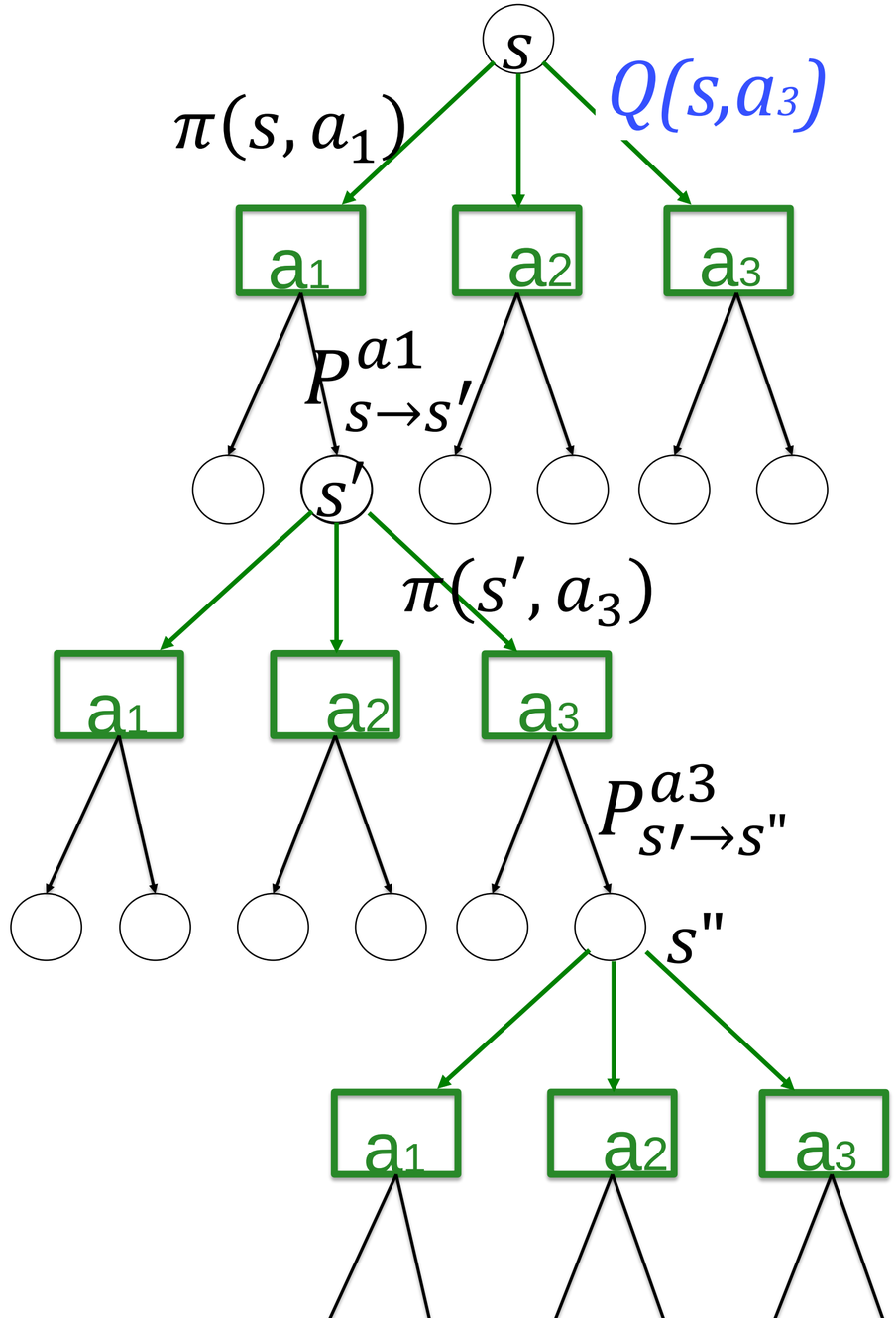
probability to choose action a in state s

$1=\sum_{a'}\pi(s,a')$

Examples of policy:
-epsilon-greedy
-softmax

Stochasticity $P^{a}_{s\to s'}$

probability to end in state s' taking action a in state s



Previous slide.

After a first action that leads to state s' starting from state s , the agent can now take a second action starting from s'.

Note that there are two different types of branching ratio:

$\pi(s, a_1)$ describes the probability that the agent uses action a1 when it is in state s – based on the agent's policy (such as epsilon-greedy)

$P_{s \to s'}^{a1}$ describes as before the probability that the agent arrives in state s' given that it chooses action a1 in state s.

As before we are interested in the expected reward. The Q value Q(s,a) describes the total accumulated reward the agent can get starting in state s with action a.

Next slide: rewards that are n steps away are discounted with a factor γ^{n}

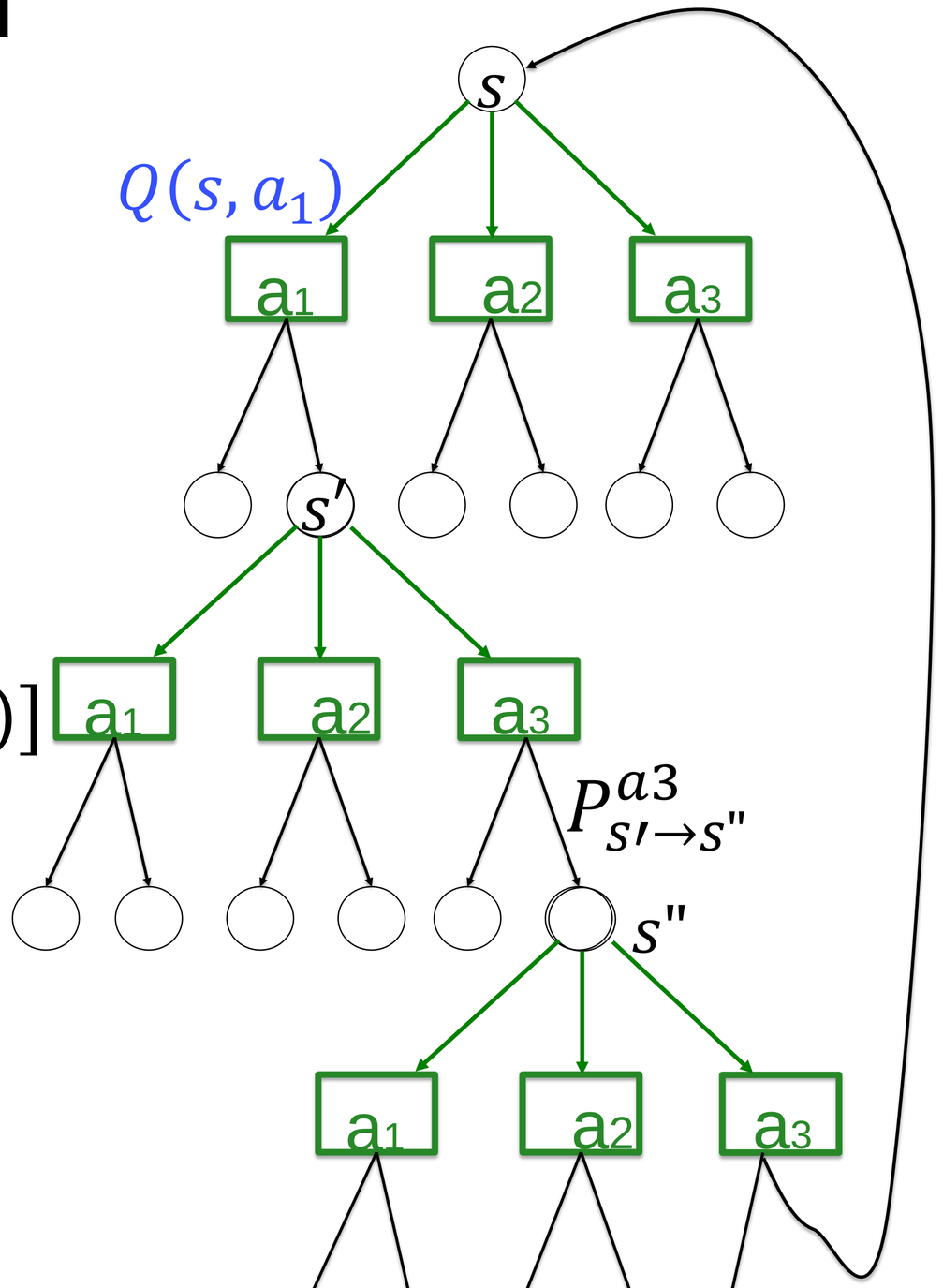
Total expected (discounted) reward

Starting in state s with action a

$Q(s,a)$ =

$$= \left\langle r_t \quad + \gamma\, r_{t+1} + \ \gamma^2 r_{t+2} + \gamma^3\, r_{t+3} + \cdots \right\rangle$$

$$= E[r_t \ + \gamma\, r_{t+1} + \ \gamma^2 r_{t+2} + \gamma^3\, r_{t+3} + \ \ldots | s, a)]$$



Discount factor: γ <1

-important for recurrent state transition graphs!

-avoids blow-up of summation

-gives less weight to reward in **far** future

Previous slide.

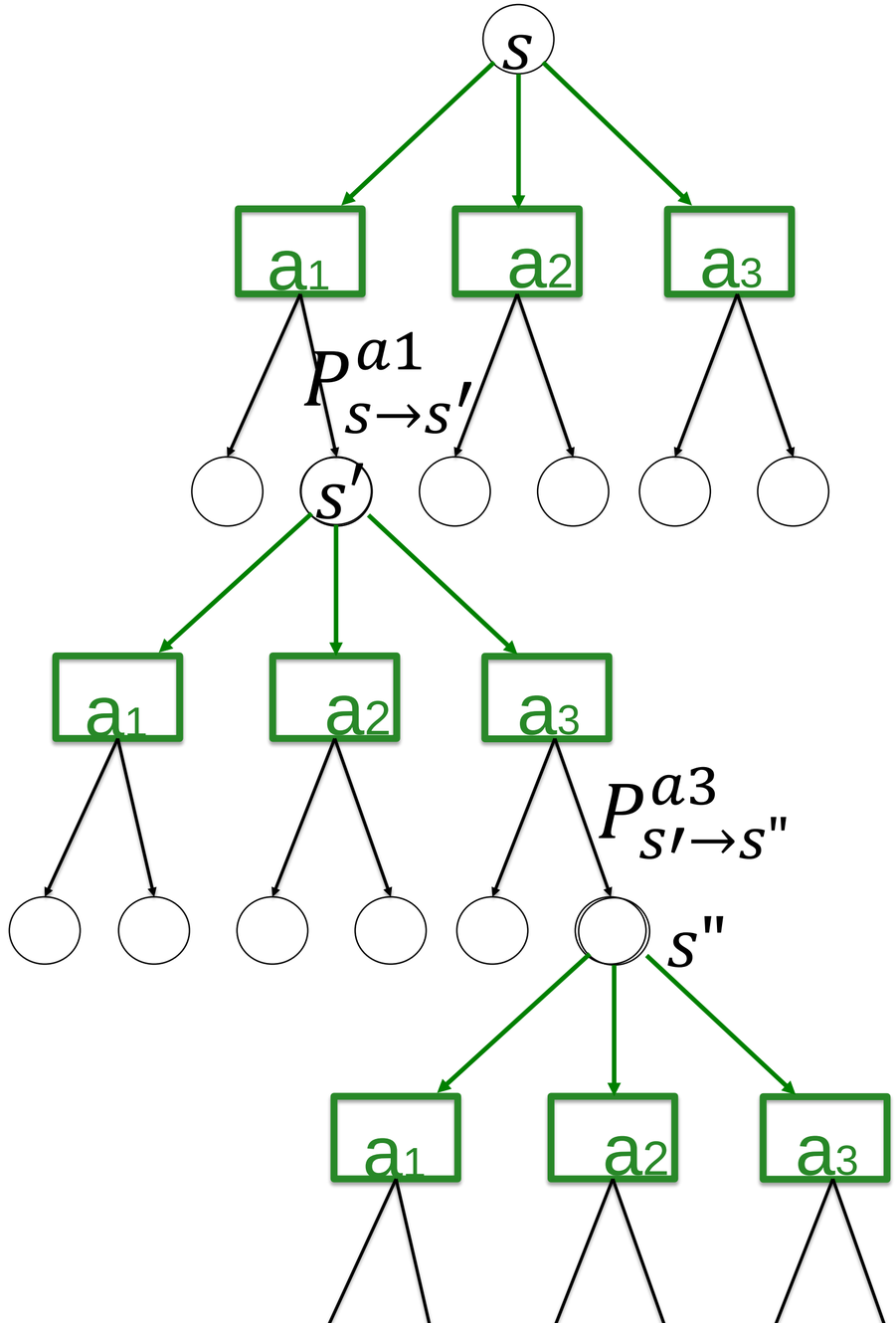
Angular brackets denote expectation (or averages over many trials, always with the same policy $\pi(s,a)$ and all starting in (s,a).

Red-font lower-case r indicates the reward collected over multiple time steps in **one single episode**, starting in state s with action a.

Expectation means that we have to take the average over all possible future paths giving each path its correct probabilistic weight.
The probabilistic weight includes the fixed policy $\pi(s,a)$
as well as the branching ratio P(s,a)

Bellman equation

Blackboard4:
Bellman eq.

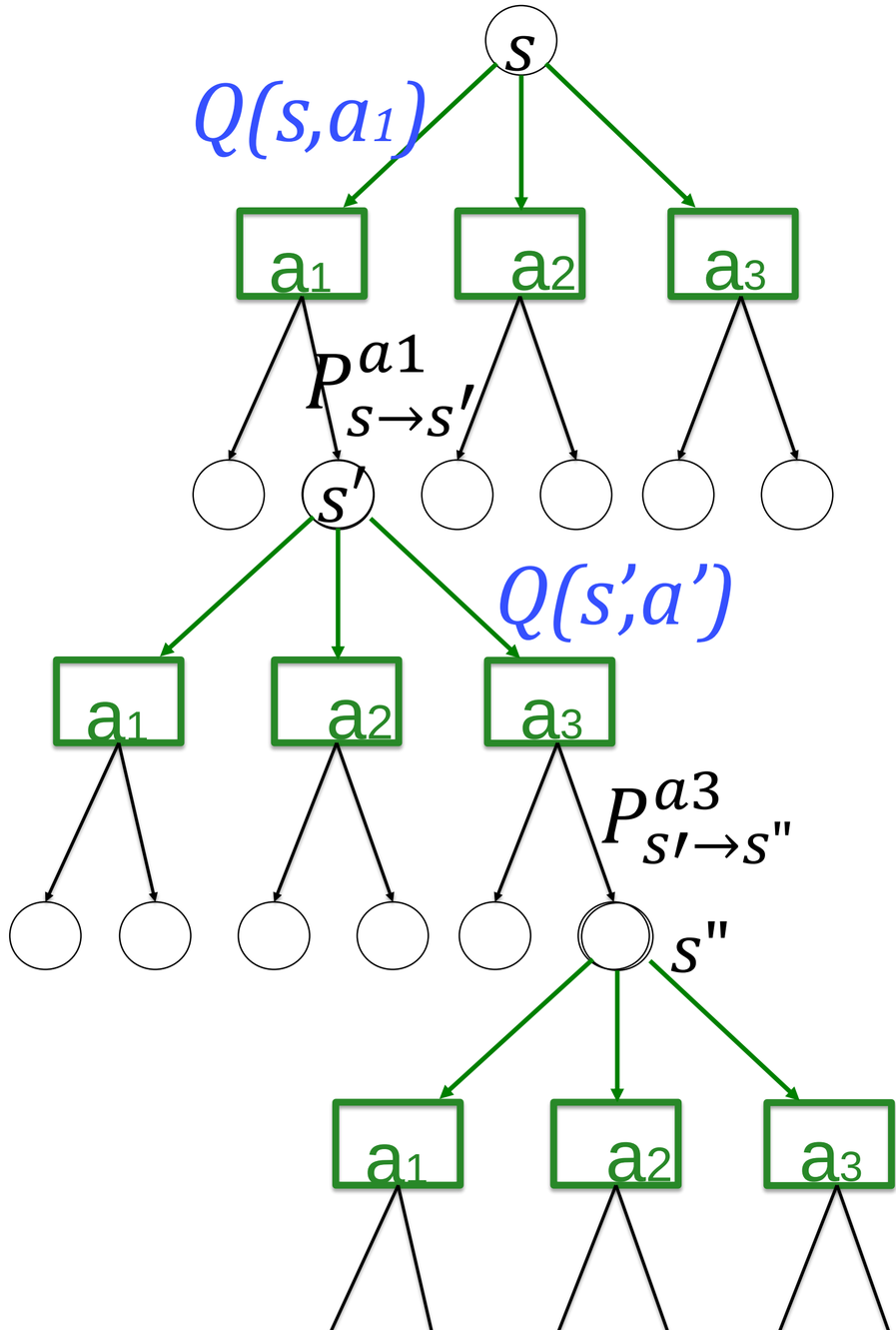


Space for calculations.

Bellman equation with policy π

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

Bellman equation = value consistency of neighboring states



Remark:

Sometimes Bellman equation is written for greedy policy: $\pi(s,a) = \delta_{a,a*}$

with action $a^* = \underset{a'}{\operatorname{argmax}}\, Q(s,a')$

Previous slide.

The Bellman equation relates the Q-value for state s and action a with the Q-values of the neighboring states.
Neighboring means reachable in a single step.
Note that the two different types of branching ratio both enter the equation.

Bottom: in the case of a greedy policy, the Bellman equation simplifies

Bellman equation (for optimal actions)

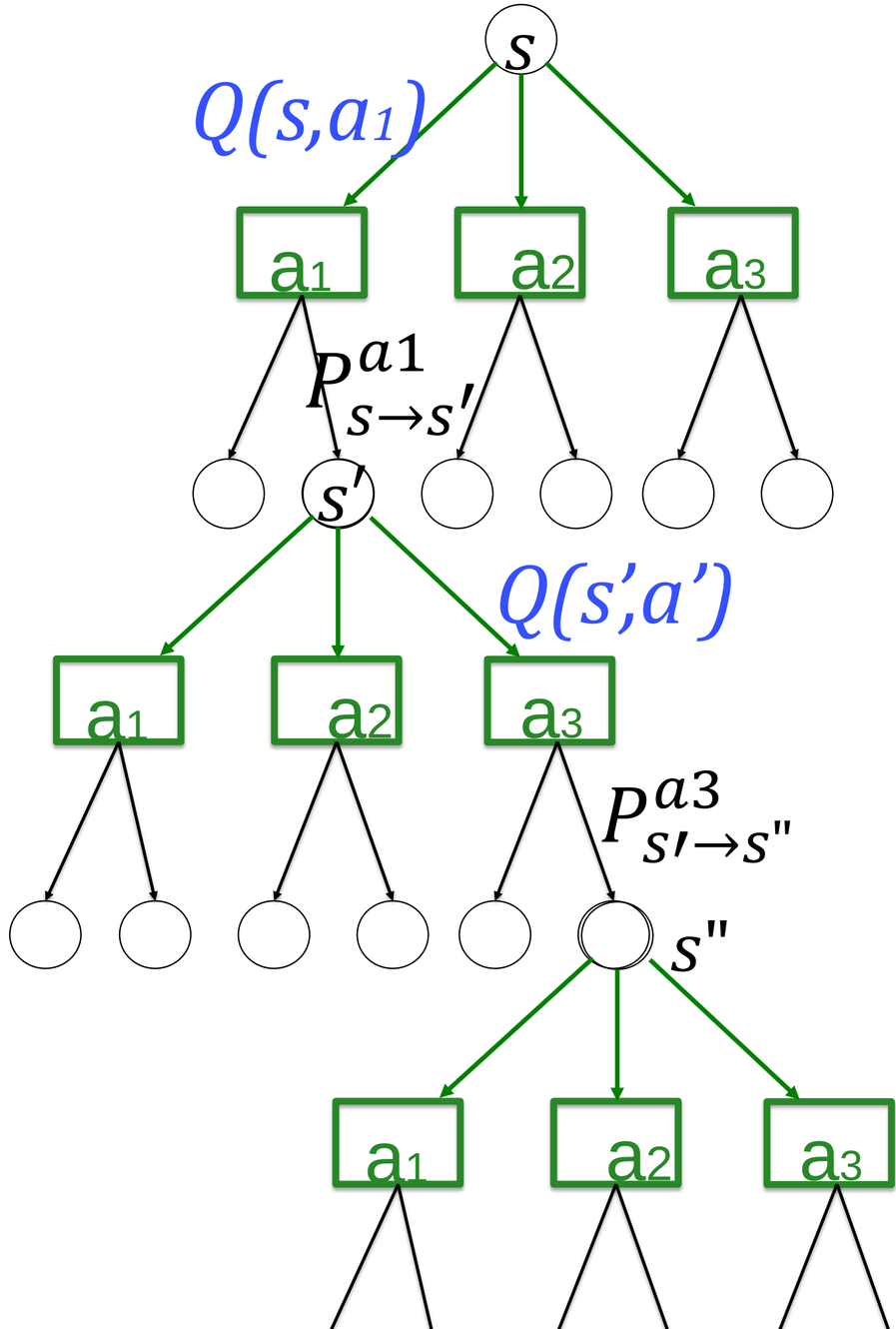
$$Q(s,a) = \sum_{s'} P^{a}_{s\to s'} \left[R^{a}_{s\to s'} + \gamma \sum_{a'} \pi(s',a') Q(s',a') \right]$$

for greedy policy:

$$\pi(s,a) = \delta_{a,a*}$$

with action $a^* = \underset{a'}{\operatorname{argmax}}\, Q(s,a')$

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



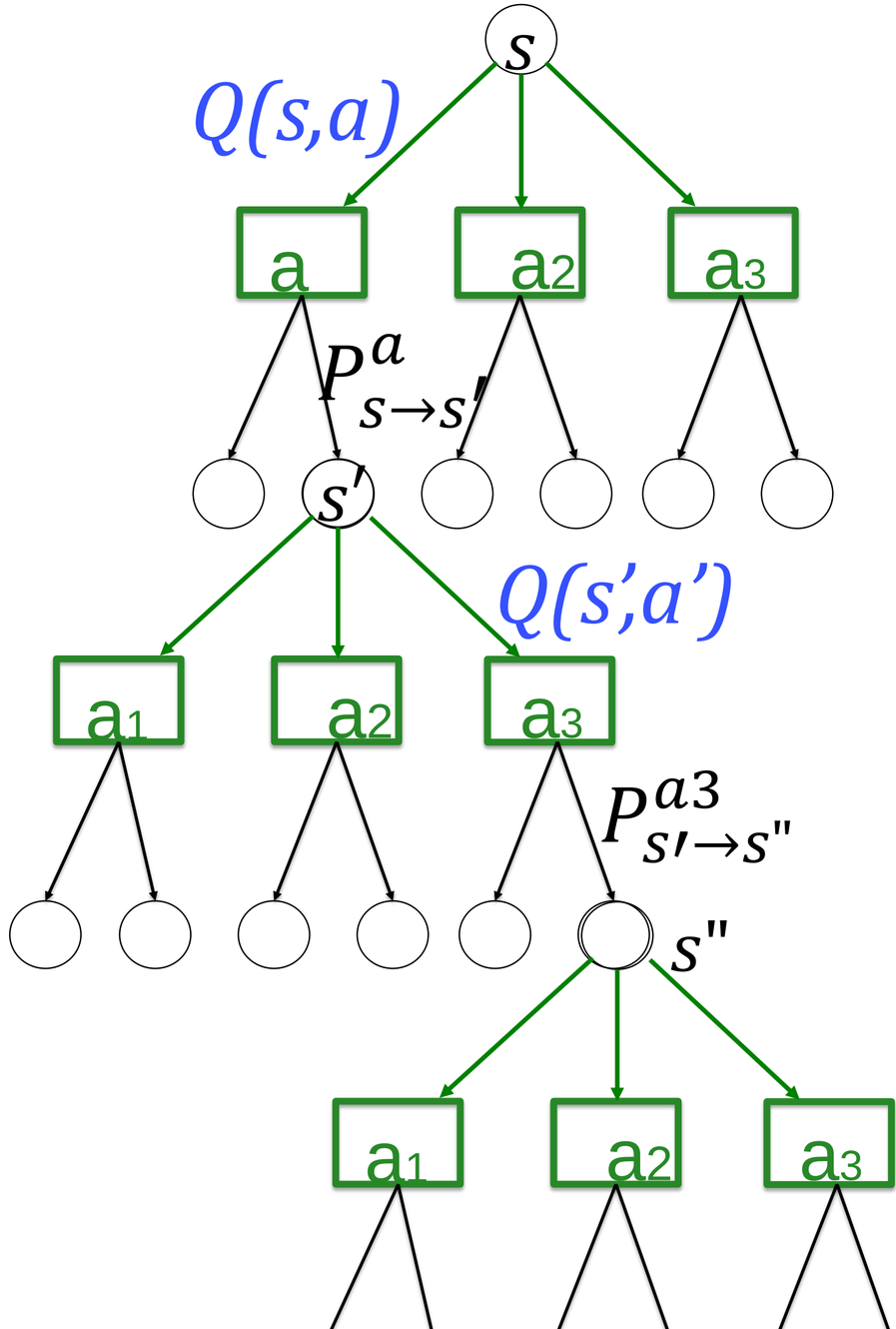
Previous slide.
For a greedy policy, the sum over actions disappears from the Bellman equation and is replaced by the max-sign.

Quiz: Bellman equation with policy π

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

[ ] The Bellman equation is linear in the variables $Q(s'a')$

[ ] The set of variables $Q(s',a')$ that solve the Bellman equation is unique and does not depend on the policy



Your comments.

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

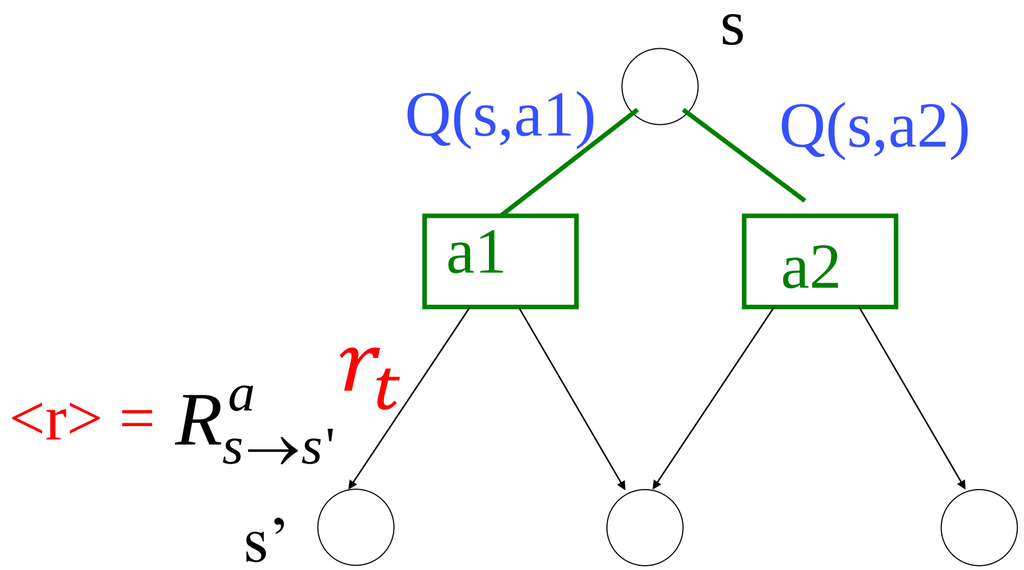
Part 6: SARSA Algorithm

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step Horizon (Bandit Problems)
- Exploration vs. Exploitation
- Bellman Equation
- **SARSA Algorithm**

Previous slide.
We not turn to the first practical algorithm, called SARSA. This is an algorithm that is widely used in the field of reinforcement learning.

Review: Iterative update of Q-values

Problem: Q-values not given



Solution: iterative update

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

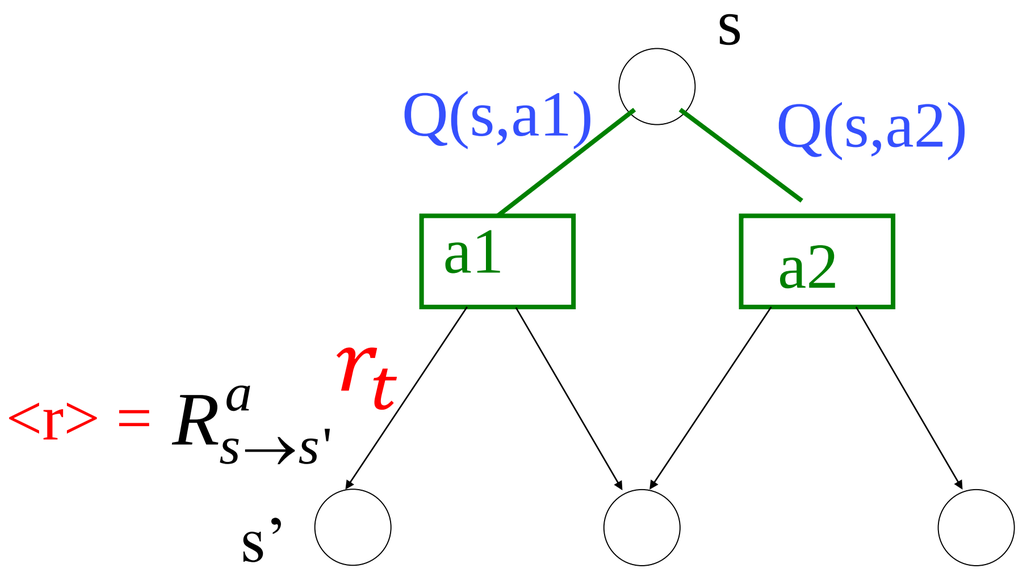
while playing with policy $\pi(s,a)$

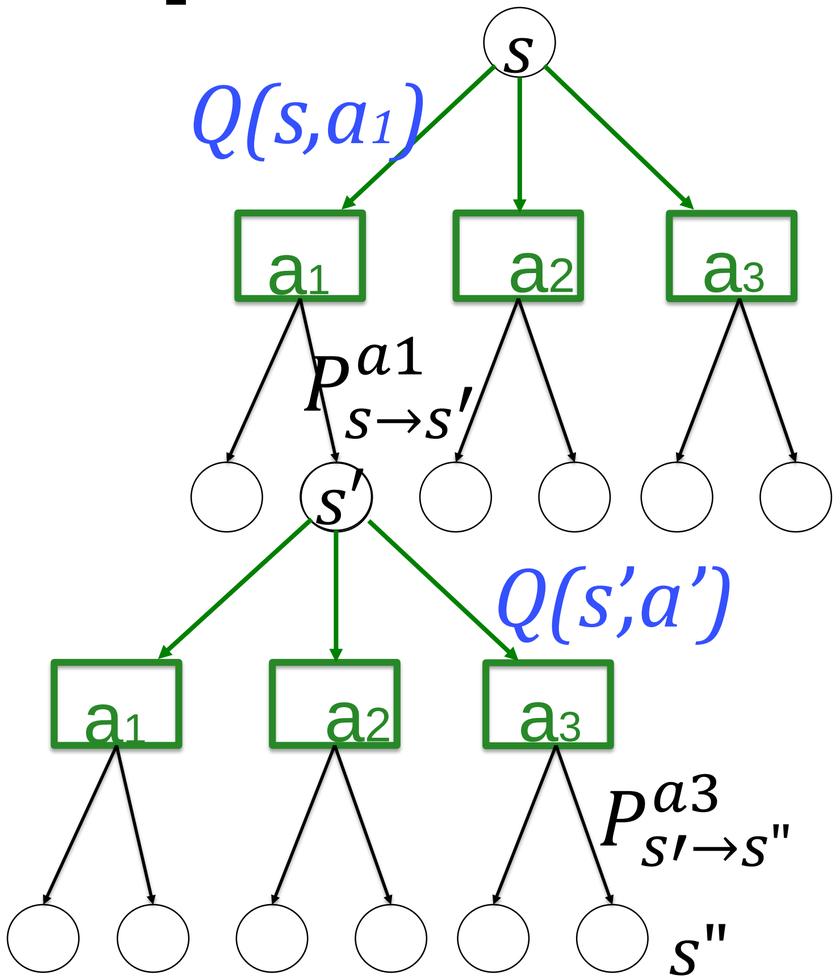
Previous slide.
Reminder: for the 1-step horizon scenario we found that we could calculate the Q-values iteratively.
We increase the Q-value by a small amount (with learning rate 0<eta<<1) if the reward observed at time t is larger than our current estimate of Q.
And we decrease the Q-value by a small amount if the reward observed at time t is smaller than our current estimate of Q.

Iterative updates with one data point at a time are also called ‘online algorithms’.
Thus our update rule is an online algorithm for the estimation of Q-values.

Iterative update of Q-values for multistep environments

Problem: Q-values not given





Solution: iterative update

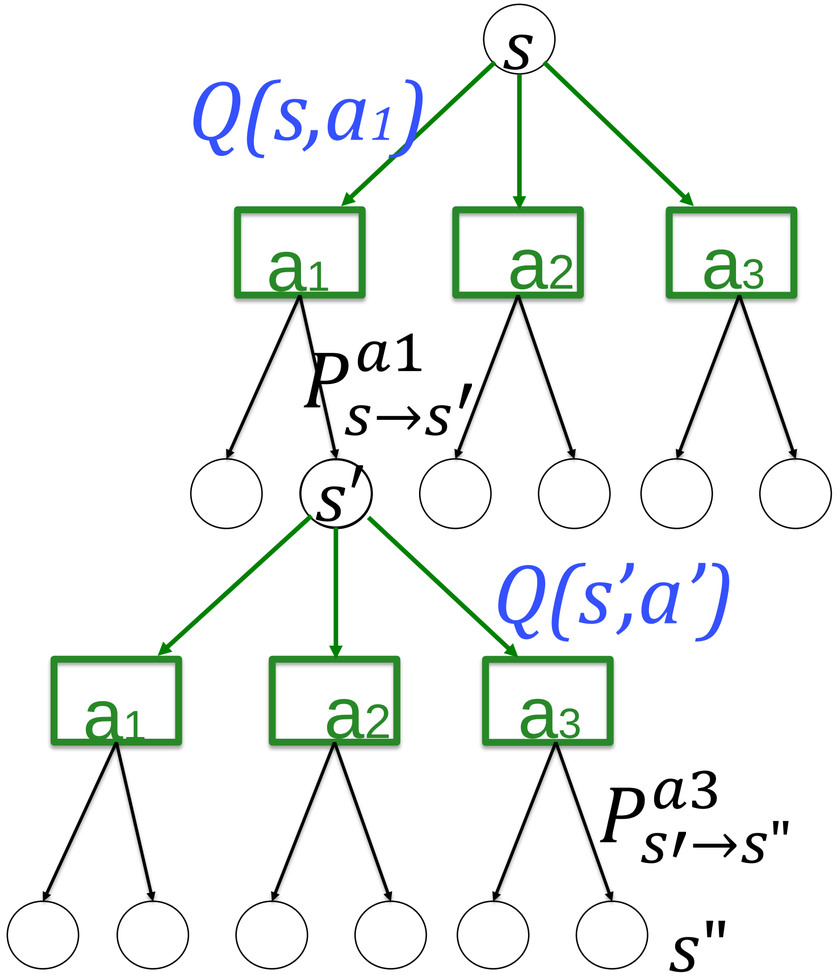
$$\Delta\hat{Q}(s,a) = \eta\,[r_t - \hat{Q}(s,a)]$$

while playing with policy $\pi(s,a)$

$$\Delta\hat{Q}(s,a) = \;?$$

Previous slide.
The question now is: can we have a similar iterative update scheme also for the multi-step horizon?

Blackboard5: SARSA update



Your notes.

Iterative update of Q-values for multistep environments

Bellman equation:

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

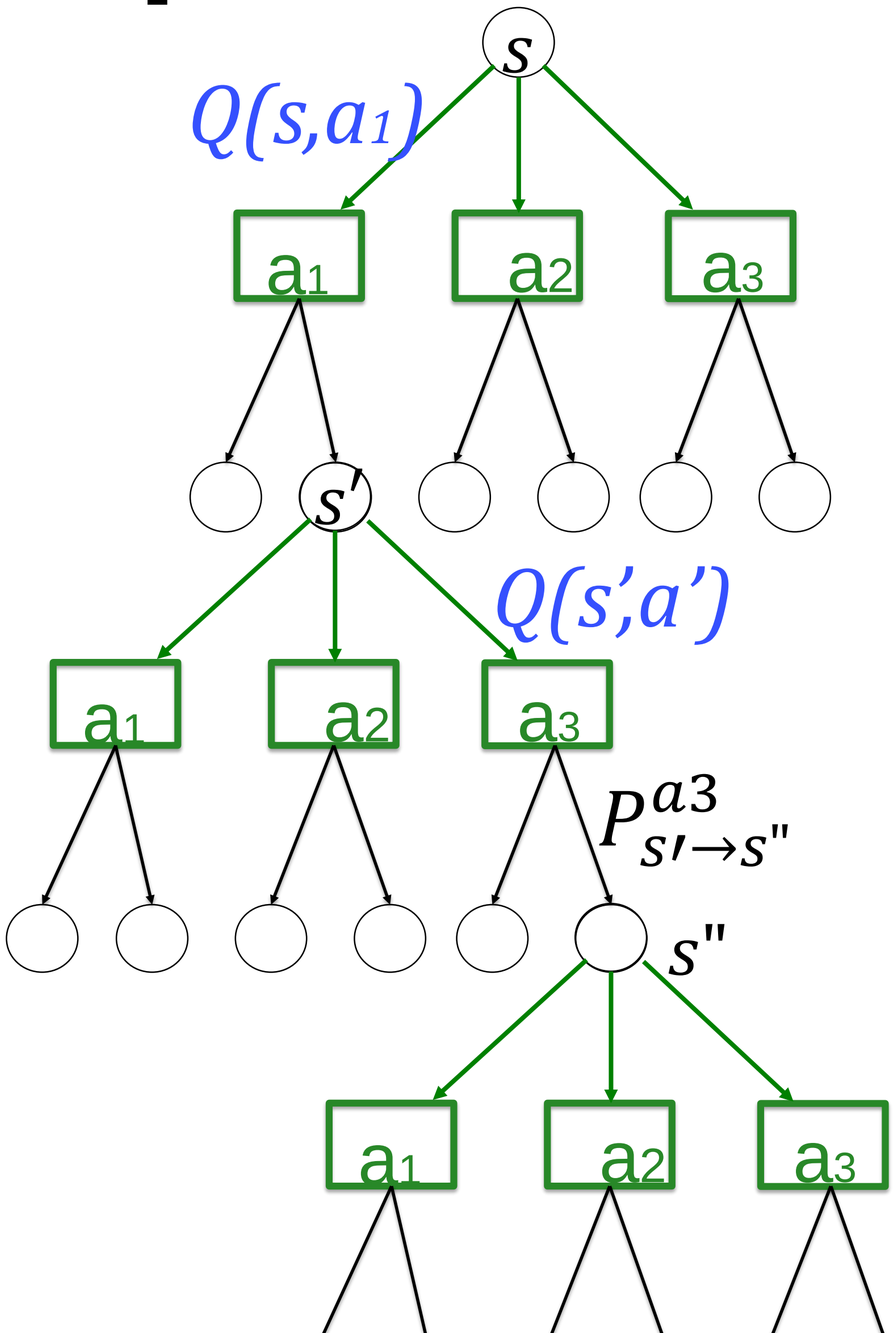
Problem:

- Q-values not given
- branching probabilities not given
- reward probabilities not given

Solution: iterative update

$$\Delta\hat{Q}(s,a) = \eta\ [r_t + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)]$$

while playing with policy $\pi(s,a)$



Previous slide.
Even for the case of the multi-step horizon, we can estimate the Q-values by an interative update:

The Q-values Q(s,a) is increased by a small amount if the sum of (reward observed at time t plus discounted Q-value in the next step) is larger than our current estimate of Q(s,a).

This iterative update gives rise to an online algorithm.

NOTE: in the following we always work with empirical estimates, and drop the ‘hat’ of the variable Q.

SARSA vs. Bellman equation

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

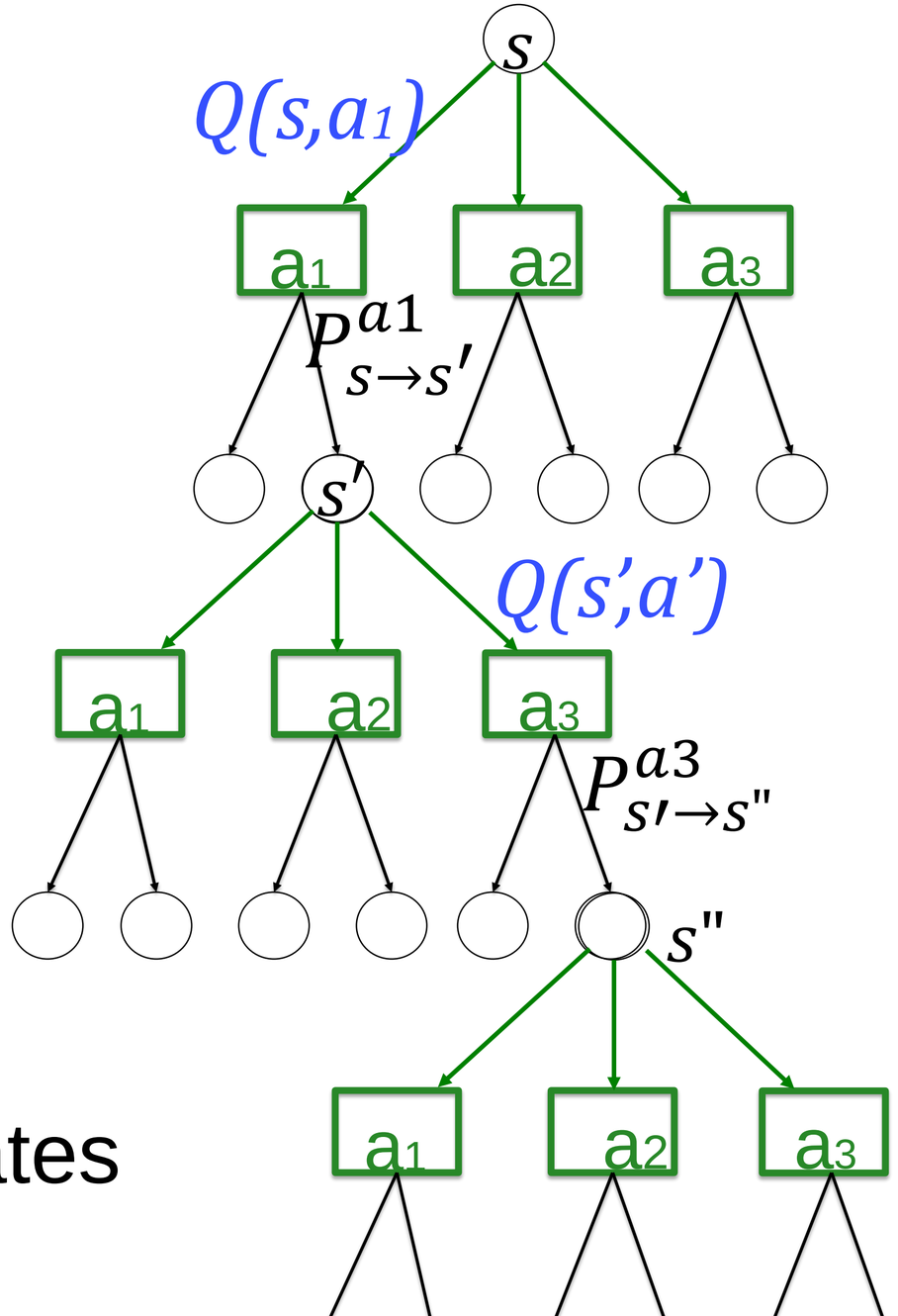
Bellman equation

= consistency of Q-values
across neighboring states

SARSA update rule

$$\Delta\hat{Q}(s,a) = \eta[r_t + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)]$$

= make Q-values of neighboring states
more consistent



Previous slide.
The Bellman equation summarizes the consistency condition:
The (average) rewards must explain the difference between Q(s,a) and γ Q(s',a') averaged over all s' and a'.
Or equivalently:
Q(s,a) must be explained by the (average) reward in the next step and the discounted Q-value in the next state.

The iterative update formula implies that Q(s,a) needs to be adapted so the current reward explains the difference between Q(s,a) and Q(s',a').

An equivalent form of writing the update is:

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

This form highlights the similarity to the case of the update rule in the cae of the one-step horizon.

SARSA algorithm

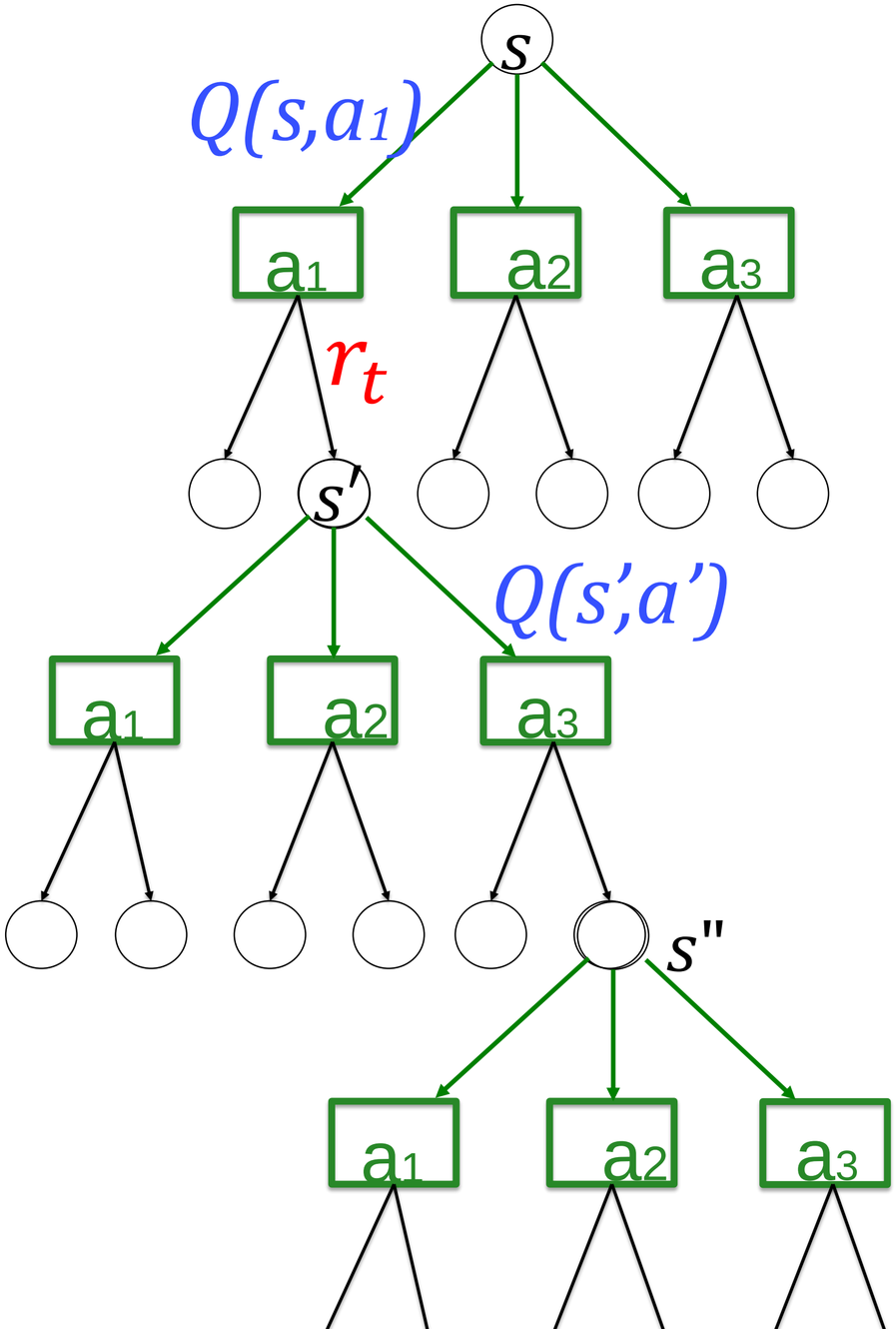
Initialise Q values
Start from initial state **s**

1) being in state **s**
   and having chosen action **a**
   [according to policy $\pi(s,a)$]
2) Observe reward **r**
   and next state **s'**
3) Choose action **a'** in state s'
   [according to policy $\pi(s,a)$]
4) Update with SARSA update rule

$$\Delta Q(s,a) = [r_t + \gamma Q(s',a') - Q(s,a)]$$

5) set: s ← s'; a ← a'
6) Goto 1)

Stop when all Q-values have converged



Previous slide.
The update rule gives immediately rise to an online algorithm. You play the game. While you run through one of the episodes you observe the state s, choose action a, observe reward r, observe next state s’ and choose next action a’. At this point in time (and not earlier) you have all the information to update the Q-value Q(s,a).

The name SARSA comes from this sequence state-action-reward-state-action.

SARSA algorithm.

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

We have initialized SARSA and played for n>2 steps.
Is the following true for the next steps?

[ ] in SARSA, updates are applied after each move.

[ ] in SARSA, the agent updates the Q-value $Q(s_{(t)},a_{(t)})$ related to the **current** state $s_{(t)}$

[ ] in SARSA, the agent updates the Q-value $Q(s_{(t-1)},a_{(t-1)})$ related to the **previous** state, once it has chosen $a_{(t)}$

[ ] in SARSA, the agent moves in the environment using the policy $\pi(s,a)$

[ ] SARSA is an online algorithm

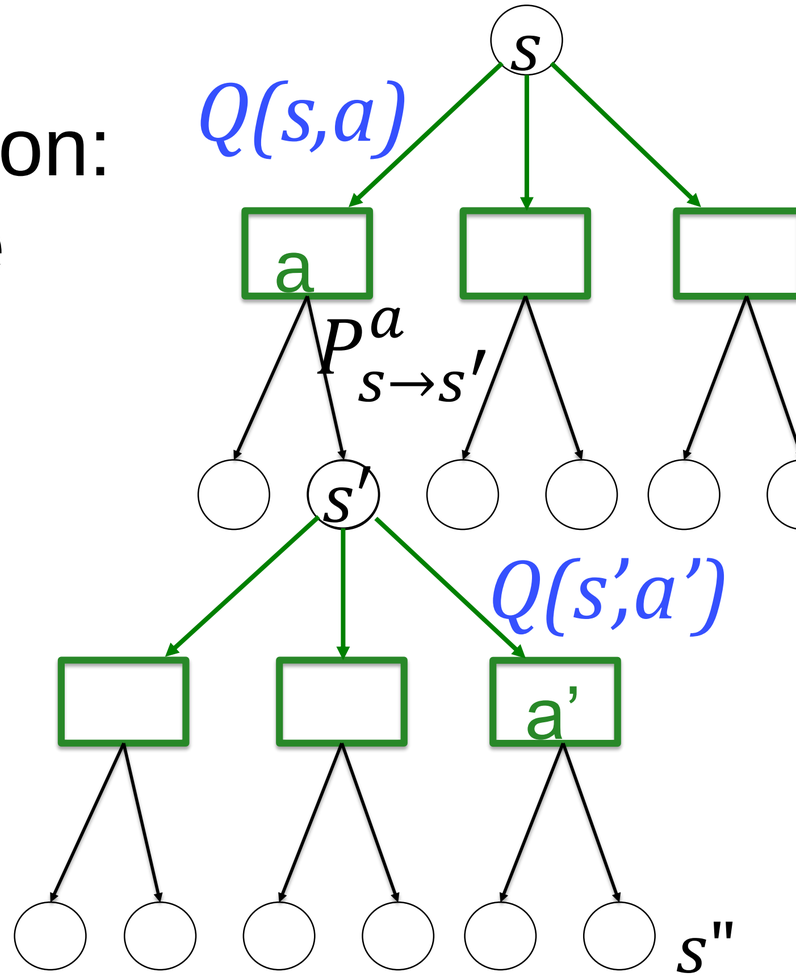
Previous slide.

Variant A: SARSA is consistent w. Bellman equation

Blackboard 6A: SARSA

We proof the following:

Suppose that we have found a set of Q-values.
We keep them frozen while evaluating expectation:
IFF $E[\Delta Q(s,a)] = 0$, then the Q-values solve the Bellman equation.



Notes:

- Expectation is taken for fixed Q-values and hence for fixed policy (consider Q-values as $\boldsymbol{\theta^{old}}$)
- Expectation $\mathrm{E}[\Delta Q(s,a)]$ is taken over all possible paths starting in (s,a). I call this ‘batch-like’.
- The length of the path is given by the needs of the update equation: Here from (s,a) to (r,s’,a’)
- Look at state-action-diagram to keep track of terms

Your comments.

Variant A: SARSA is consistent w. Bellman equation

Blackboard 6A: SARSA

We have proven the following:

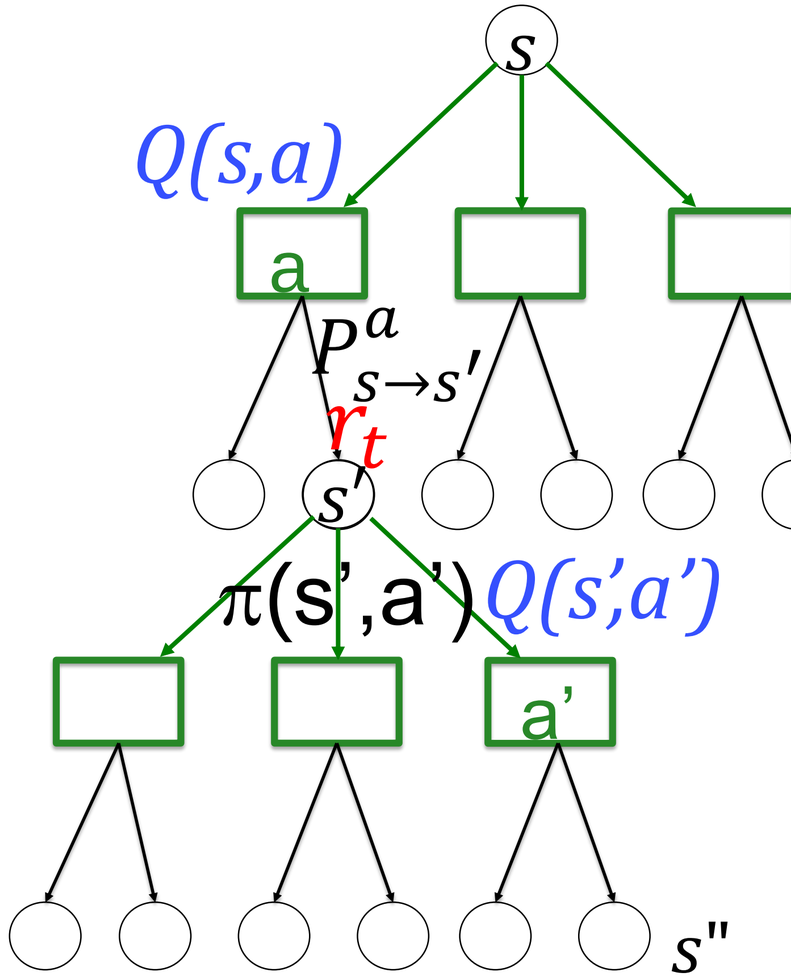
Suppose that we have found a set of Q-values.

We keep them frozen while evaluating expectation.

IFF $E[\Delta Q(s,a)] = 0$, then the Q-values solve the Bellman equation.

$$E[\Delta Q(s,a)] = \eta E[r_t + \gamma Q(s',a') - Q(s,a)]=0$$

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



Look at graph to take expectations:

- if algo is on a branch (s,a), all remaining expectations are “given s and a”

Previous slide.
This is version A of the theorem. In the proof, we exploit the condition:

$$E[\Delta Q(s,a)|s,a] = 0$$

In order to take the expectations, we look at graph:
- if in the evaluation we are in state s', all remaining expectations are "given s'"
- if we are on a branch (s,a), all remaining exp. are "given s and a".

We exploit that all Q-values and the policy are fixed while we evaluate the expectation. Hence E[Q(s,a)] = Q(s,a).

Note that the proof works in both direction. If the Q-values are those of the Bellman equation, the expected SARSA update step vanishes. And if the expected SARSA update step is zero, then the Q-values correspond to the Bellman equation. Both direction work under the assumption of a fixed policy.

The stronger theorem (with fluctuations, version B) is sketched in the appendix (and video).

Additional Notes: This weaker theorem (variant A that corresponds to the one on the previous slide) takes expectations for FIXED Q-values. We can interpret these expectations as the following 'batch' computation

We assume a fixed policy (i.e., under the assumption of a fixed set of Q-values) and a 'batch version' of SARSA. Batch-SARSA means that in order to evaluate $E[\Delta Q(s,a)|s,a]$ we use a large number of starts from the same value (s,a) each time running one step up to (s',a') [note that this gives different (s',a'). Once the number of starts is large enough to get a full sample of the statistics we update Q(s,a). If the updates with the batch-SARSA do not lead to a change of Q values (for all state-action pairs), then this means that batch-SARSA has converged to the Bellman equation for this fixed policy. (That was the theorem in the main text).

Batch-SARSA is a computational implementation of the way many statistical convergence proofs work: you assume that you average over a full statistical sample of all possibilities given your current state or the current state-action pair. Expectation signs in the update step imply updating over a 'full batch of data'. In this approach Q-values no longer fluctuate, and hence do not need expectation signs; the policy no longer fluctuates and also does not need expectations signs.

Your comments.

Variant B: Bellman equation and SARSA: theorem for small η

Setting: 'temporal averaging'

The SARSA algo has been applied for a very long time, using updates

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

IF (i) averaged over many update steps

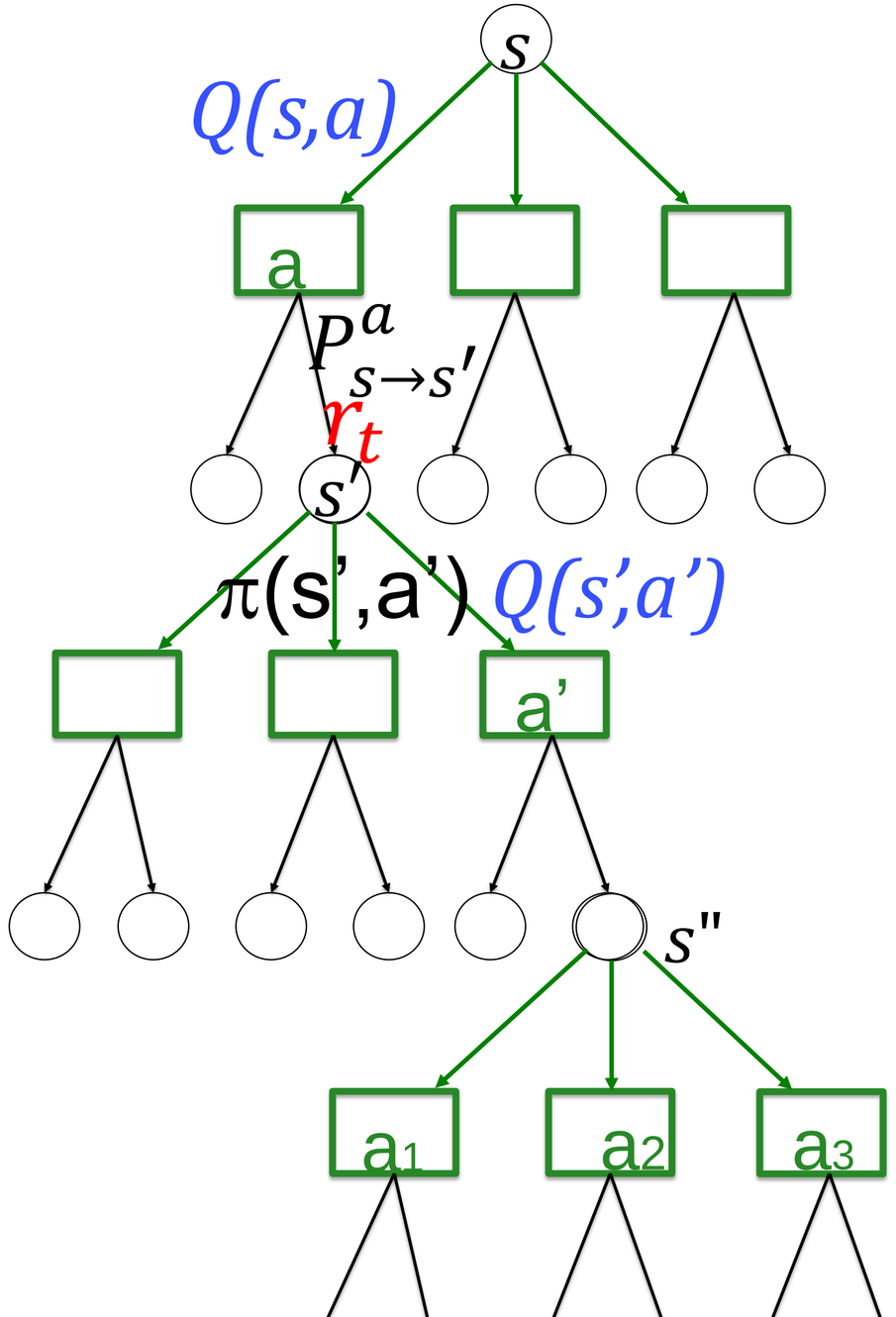
$$\langle \Delta Q(s,a) | s,a \rangle = 0$$

(ii) learning rate η is small $(\eta \to 0)$

THEN: fluctuations of Q are negligible and the set of expected Q-values solves the Bellman eq.

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

with the current policy $\pi(s',a')$



Previous slide. There are two different versions of the theorem. This version B is proven in the annex, in a fashion similar to the case of the 1-step horizon.
In class (earlier slides) we have shown for the multi-step horizon a weaker statement: **Expections over SARSA updates are consistent with the Bellman equation** if the expected update vanishes: **In version A, we assume that Q-values are fixed (frozen) when we take the expectation.**

Note that taking the expectation in version A is different from averaging over update steps in version B. In version A, taking the expectation means that we average over all possible outcomes in the **current** situation, with momentarily fixed Q-values and **fixed policy** (i.e., the one induced by the set of Q-values at time t). This distinction is important, because **for a fixed policy averaging is relatively easy.**

However, when averaging over time steps as in variant B, the Q-values and policy are different in each time step, and the proof therefore requires a limit $\eta \rightarrow 0$, so that changes can be neglected.
Hence there are therefore two versions of the theorem and two proof-sketches:
Blackboard 6A. On the earlier slides, we assume Q-values are fixed and do not fluctuate.
Blackboard 6B. In the Annex, we assume that Q-values may fluctuate slightly round their stable values. This approach gives additional insights into the situation of the online SARSA, once it has converged in expectation.

Teaching monitoring – monitoring of understanding

[ ] today, up to here, at least 60% of material was new to me.

[ ] up to here, I have the feeling that I have been able to follow (at least) 80% of the lecture.

Summary: Reinforcement Learning and SARSA

Now:
Exercise session

Learning outcome and conclusions:

- **Reinforcement Learning is learning by rewards**
  → world is full of rewards (but not full of labels)
- **Agents and actions**
  → agent learns by interacting with the environment
  → state s, action a, reward r
- **Exploration vs Exploitation**
  → optimal actions are easy if we know reward probabilities
  → since we don't know the probabilities we need to explore
- **Bellman equation**
  → self-consistency condition for Q-values
- **SARSA algorithm:** state-action-reward-state-action
  → update while exploring environment with current policy

Reinforcement Learning Lecture 1
Reinforcement Learning and SARSA

Wulfram Gerstner
EPFL, Lausanne, Switzerland

Part 6B: SARSA is consistent with Bellman equation

- Examples of Reward-based Learning
- Elements of Reinforcement Learning
- One-step Horizon (Bandit Problems)
- Exploration vs. Exploitation
- Bellman Equation
- **SARSA Algorithm is consistent with Bellman equation**

Exercise 4: SARSA for Linear Track.

Exercise 4 (at 15h15)

Exercise 4. SARSA algorithm

In the lecture, we introduced the SARSA (state-action-reward-state-action) algorithm, which (for discount factor $\gamma = 1$) is defined by the update rule

$$\Delta Q(s,a) = \eta \left[r - \left(Q(s,a) - Q(s',a') \right) \right] , \tag{1}$$

where s' and a' are the state and action subsequent to s and a. In this exercise, we apply a greedy policy, i.e., at each time step, the action chosen is the one with maximal expected reward, i.e.,

$$a_t^* = \arg\max_a Q_a(s,a) . \tag{2}$$

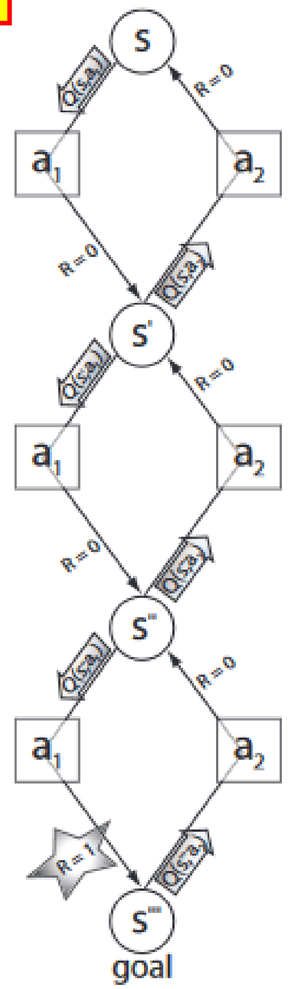
If the available actions have the same Q-value, we take both actions with probability 0.5.

Consider a rat navigating in a 1-armed maze (=linear track). The rat is initially placed at the upper end of the maze (state s), with a food reward at the other end. This can be modeled as a one-dimensional sequence of states with a unique reward ($r = 1$) as the goal is reached. For each state, the possible actions are going up or going down (Fig. 2). When the goal is reached, the trial is over, and the rat is picked up by the experimentalist and placed back in the initial position s and the exploration starts again.

a. Suppose we discretize the linear track by 6 states, $s_1, \ldots, s_6$. Initialize all the Q-values at zero. How do the Q-values develop as the rat walks down the maze in the first trial?

b. Calculate the Q-values after 3 complete trials. How many Q-value values are non-zero? How many trial do we need so that information about the reward has arrived in the state just 'below' the starting state?

c. What happens to the learning speed if the number of states increases from 6 to 12? How many Q-values are non-zero after 3 trials? How many trial do we need so that information about the reward has arrived in the state just 'below' the starting state?



ure 2: A linear maze.

Annex: Variant B - SARSA and Bellman equation (proof for small η)

Setting: 'Temporal Averaging'

The SARSA algo with stochastic policy π has been applied for a long time with updates

$$\Delta\hat{Q}(s,a) = \eta\,[r_t + \gamma\hat{Q}(s',a') - \hat{Q}(s,a)]$$

IF (i) learning rate η is small; AND IF

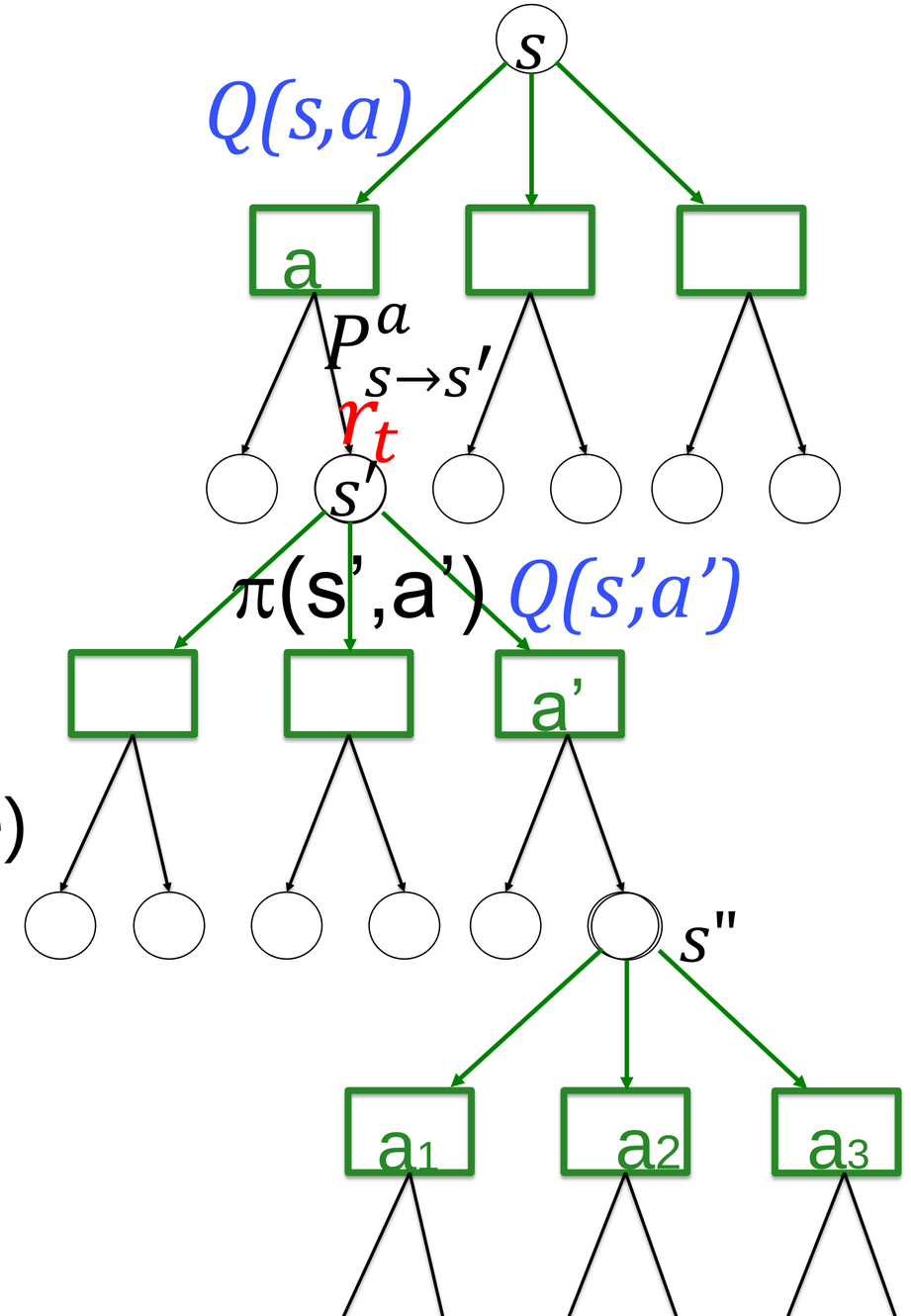
(ii) for all Q-values

$$\langle \Delta Q(s,a) | s,a \rangle = 0$$

THEN the **expectation** values (temporal average) of the set of $\hat{Q}$-values solves the Bellman eq.

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

with the current policy $\pi(s',a')$



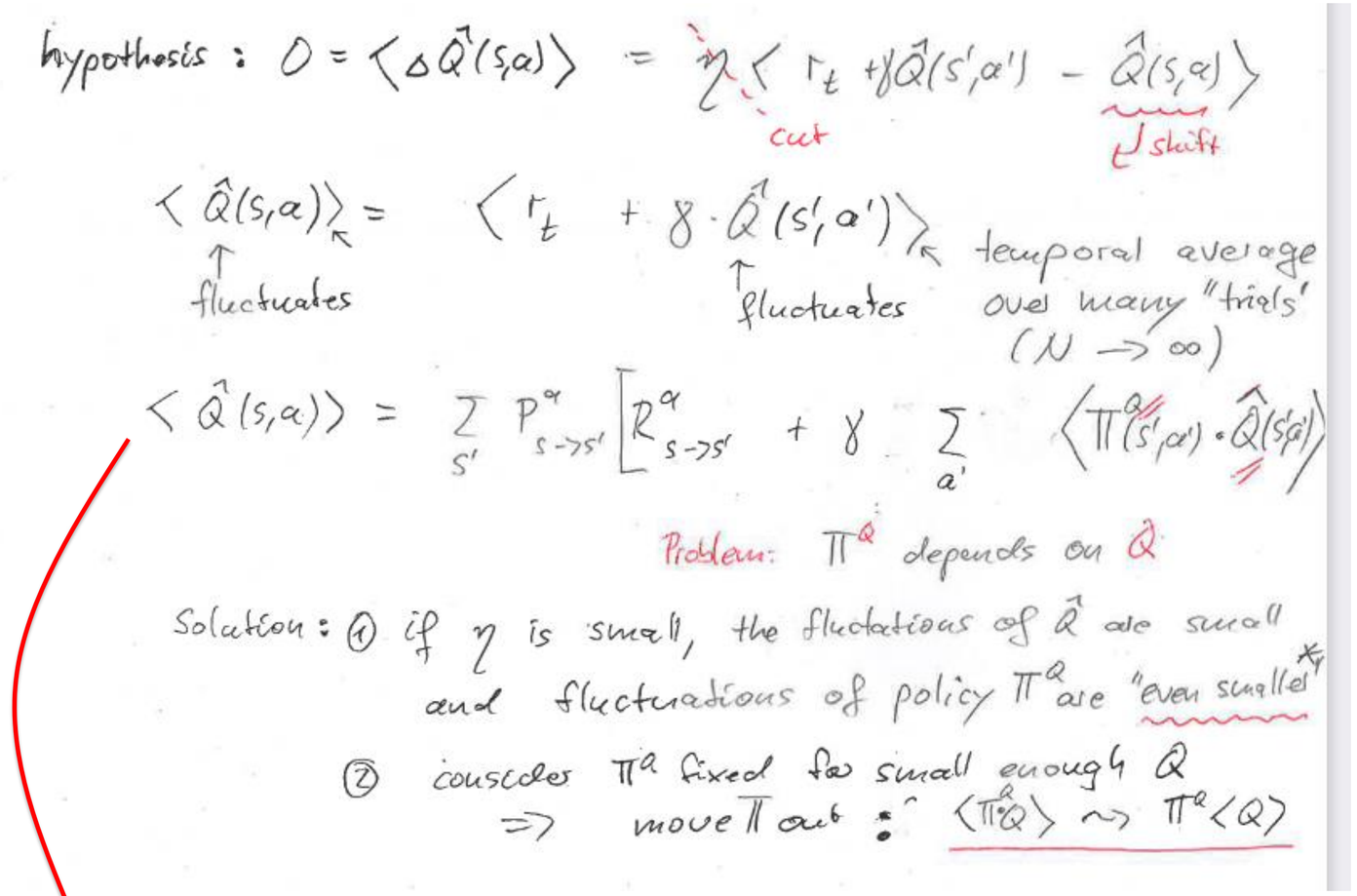
Notes: A few points should be stressed:

1. This is not a convergence theorem. We just show consistency as follows:
   if SARSA has converged then it has converged to a solution of the Bellman equation.
2. In fact, for any finite η the SARSA Q-values (Q-hat) fluctuate a little bit. It
   is only the EXPECTATION value of the Q-hat which converges.
3. We should keep in mind that SARSA is an on-policy online algorithm for arbitrary
   state-transition graphs. Hence the value Q-hat at (s,a) and (s'a') will both fluctuate!
4. The policy depends on these Q-hat-values and hence fluctuates as well.
   To keep fluctuations of the policy small, we need small η .
   We imagine that all Q-hat values fluctuate around their expectation value
   with small standard deviation. As a result, π also fluctuates around a 'standard' policy.
5. The fluctuations of the policy can be smaller than that of the Q-values: for
   example in epsilon-greedy, you first order actions by the value of Q(s,a),
   and then only the rank of Q(s,a) matters, not their exact values. In the proof we assume
   that the fluctuations of the policy become negligible (→ shift policy outside expectation).
6. We show that the Q-values in the sense of Bellman are the expectation values of
   the Q-hat in the sense of SARSA.
7. Expectations are over many trials of the ONLINE SARSA.

The statement and proof is different to slide 127 and to the book of Sutton and Barto.

Variant B – temporal averaging: SARSA is consistent with Bellman (proof for small η)

Claim for Online SARSA: $\hat{Q}(s,a)$ jitters, but its mean $\langle \hat{Q}(s,a) \rangle$ solves Bellman



Look at graph to take expectations over update steps

- if algo is in state s, expectations "given s"
- if algo is on a branch (s,a), all remaining expectations are "given s and a"

*example of π^Q "even smaller": ε- greedy
only rank-order of Q-values matters: best / 2nd best / ...
if fluctuations $|\Delta \hat{Q}(s',a')| \ll Q(s',\text{best}) - Q(s',\text{2nd best})$
then π^Q remains stable!

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

solves Bellman!

Notes: This is a proof sketch of the consistency of online SARSA **(Variant B).** We allow all Q-values to fluctuate around their expectation (visualized as ‘temporal averaging’) , but we still have to keep fluctuations of the policy negligibly small.

If we allow for small fluctuations of the policy, then we have to realize that these fluctuations are correlated with the fluctuations of Q-values. Thus the evaluation of the product E(π Q) is not trivial. Moreover, correlations can lead to a shift of the value and make the result inconsistent with the Bellman equation.
This variant B therefore only works in the limit of vanishing learning rate.

The other theorem **(Variant A** that corresponds to the one on the slides in the main part of this lecture) takes expectations for FIXED Q-values. We can interpret these expectations as corresponding to a ‘batch’ computation.
Batch-SARSA is a computational implementation of the way many statistical convergence proofs work: you assume that you average over a full statistical sample of all possibilities given your current state or the current state-action pair. Expectation signs in the update step imply updating over a ‘full batch of data’. In this approach Q-values no longer fluctuate, and hence do not need expectation signs; the policy no longer fluctuates and also does not need expectations signs.