

Game Theory & Reinforcement Learning

Felix Putze

12.7.2012

Lecture „Cognitive Modeling“
SS 2012

Modeling Decision Behavior

- To predict the actions of a human (e.g. the user of a system), we need to model how he decides for a certain behavior
- Steps of decision making (according to Kepner & Tregoe, 1965):
 1. Establish objectives
 2. Priorize objectives
 3. Establish available actions
 4. Evaluate actions against objectives
 5. Make a tentative decision for an action
 6. Evaluate side-effects of chosen action
- → Many alternatives to design the evaluation and prioritization steps of decision making

Homo Economicus

- A main assumption of most formal models of decision making is the paradigm of the Homo oeconomicus (Mill, 1870ies):
 - Self-interested (in contrast to deciding for or against others)
 - Rational: Makes decisions with maximized utility
- Well suited for modeling of decision making
 - Find mathematical models of utility
 - Determine algorithms to maximize utility
- Critics of this model:
 - Bounded rationality: Model assumes too much economic knowledge and forecasting ability of the agent
 - Irrationality: Agents do not behave rational (e.g. risk averse behavior, inner conflicts between short-term and long-term goals)
 - Motivation: Not only extrinsic motivation but also intrinsic motivation determines decisions (e.g. altruistic self-sacrifice)

Utility Functions

- Von Neumann, Morgenstern (1944): Utility theory
 - B = set of baskets for consumption
 - $u: B \rightarrow \mathbb{R}$ maps each basket to a real rating
 - $u(X) > u(Y)$ iff the agent prefers X over Y ($X \succ Y$)
 - Not for every set of preferences, a utility function exists
- The Von-Neumann-Morgenstern utility theorem states that for every agent that is *rational*, a utility function exists:
 - Completeness: For every pair X, Y , we have: $X \succ Y$ or $Y \succ X$ or $X = Y$
 - Transitivity: If $X \succ Y$ and $Y \succ Z$ then : $X \succ Z$
 - Continuity: If $X \succ Y \succ Z$ then there is a probability p that $pX + (1-p)Z = Y$
 - Independence: If $X \succ Y$, for any Z and probability p : $pX + (1-p)Z \succ pY + (1-p)Z$
- u allows to quantify preference in a way which allows to add utilities , e.g. to calculate expected utility in a lottery
 - Note that we still have in general $u(x) + u(y) \neq u(x \text{ and } y)$

Violations of Rationality Assumptions

- While the assumptions which guarantee the existence of a utility function sound modest, there is evidence for violations
- Example 1: Violation of Transitivity
 - Pairwise comparison of different lotteries showed intransitive preference relations caused by different evaluation strategies (Tversky, 1969)
- Example 2: Violation of Independence
 - Allais-Paradox (1953): Compare the following choices between two gambles:

A	Probability	Payoff
Gamble 1	1.0	\$ 500,000
Gamble 2	0.1	\$ 2,500,000
	0.89	\$ 500,000
	0.01	\$ 0

B	Probability	Payoff
Gamble 1	0.11	\$ 500,000
	0.89	\$ 0
Gamble 2	0.1	\$ 2,500,000
	0.9	\$ 0

- For choice A, people prefer Gamble 1, for choice B, people prefer Gamble 2
- However, between A and B we removed the same probability of 0.89 for winning \$ 500,000 for both gambles

Example for „sub-optimal“ Utility Function

- If we model decision making as rational optimization of a utility function, how can we model “irrational” behavior?
- Sankt-Petersburg paradox: How much would you pay for participation in the following (repeatable) experiment?
 - Flip a coin until „tail“ comes up for the first time
 - Payoff: 2^k €, if the coin was flipped k times
- Expected value = $\infty \rightarrow$ Accepting any price to play this game seems to be the “rational” decision
- This result is neither intuitive nor in accordance with empirical evidence on human behavior
- Possible solution: diminishing utility, e.g. utility depends in a non-linear, capped way on the received payoff: $u(x) = 1 - e^{-x}$



Example: Prisoners' Dilemma (PD)

- Two bank robbers A and B are captured for speeding, locked in separated cells and confronted with the following choices:
 - If both confess the crime, they are locked in for 5 years
 - If both don't confess, they are locked in for 1 year for minor accusations
 - If one confesses and the other does not, the first is set free as principal witness and the other is locked in for 10 years
- We can display the criminals' utility function (proportional to the prison sentence) in a table:

	A: confess	A: don't confess (cooperate)
B: confess	$U_A = -5, U_B = -5$	$U_A = -10, U_B = 0$
B: don't confess (cooperate)	$U_A = 0, U_B = -10$	$U_A = -1, U_B = -1$

Game Theory

- Model moments of decisions as *game*:
 - Multiple players have a set of actions A_i and a utility function U_i
 $A_1 \times \dots \times A_n \rightarrow R$ (i.e. utility depends on actions of all players)
 - All players simultaneous decide for one action
 - All action choices are revealed simultaneously and utility is received
- Each player decides for a *strategy* which describes which action to take
 - Pure strategy: Deterministically chose one specific action
 - Mixed strategy: Probability distribution across all available actions
- Prisoner's dilemma can be directly transformed to a game
 - $A_1 = A_2 = \{ \text{confess, don't confess} \}$
 - $U_1 = U_2 = \dots$ (see table on previous slide)
 - This matrix is often called the normal form of a game
- Game theory is used to model decision situations from stock market over marriage proposals to decisions of peace and war

- | | | | | | | | | |
|------------------|----------------------|----------------------|---|----------------------|----------------------|---|------------|----------------------|
| | A: confess | A: don't confess | $-5 > -10$

$0 > -1$ | | A: confess | $-5 > -10$

$0 > -1$ | | A: confess |
| B: confess | $U_A = -5, U_B = -5$ | $U_A = -10, U_B = 0$ | | B: confess | $U_A = -5, U_B = -5$ | | B: confess | $U_A = -5, U_B = -5$ |
| B: don't confess | $U_A = 0, U_B = -10$ | $U_A = -1, U_B = -1$ | B: don't confess | $U_A = 0, U_B = -10$ | | | | |

Nash Equilibrium

- Nash equilibrium: A set of strategies for every player for which holds: No individual player can improve its utility value by changing only his action
 - Nash equilibrium is a natural simultaneous endpoint of a local optimization process for each player
 - Each set of strategies which is reached through elimination strategy is a Nash equilibrium
- Every two player game has at least one Nash equilibrium
 - Not every Nash equilibrium can be reached using elimination strategy
 - Some games contain more than one Nash equilibrium
 - The employed strategies may be mixed (e.g. non-deterministic)
 - If the equilibrium is unique, we can assume that rational players will always chose the corresponding actions

Examples of Advanced Game Theory

- Extension 1: Sequential (dynamic) games
 - Not every game involves a single simultaneous decision
 - Sometimes, players can react to the others' actions
 - Represent game as tree containing sub-games at every node
 - Strategy now comprises an action in every sub-game of the tree
 - Theory developed by Prof. Reinhard Selten (only German Nobel laureate for economy)
 - Introduces possibility of (credible and incredible) threats, i.e. actions which are never executed but still influence the other players' decisions
- Extension 2: No perfect knowledge
 - Basic theory assumes perfect knowledge of all players' utility functions
 - Sometimes, we do not know the *type* of player (e.g. which outcomes he does prefer)
 - However, we can send and receive signals about a player's type (e.g. by selecting certain actions in a game with multiple decisions)
 - Use Bayesian model to represent uncertainty about players' types

Empirical Validation

- Game theory as a model of human behavior can be evaluated empirically in the real world and in controlled experiments
- Many real world applications of prisoners' dilemma:
 - Environment: Control of CO₂ emissions by regulation of industry
 - Sports: Taking illegal steroids
 - Economy: Aggressive marketing on a market with a fixed number of customers (e.g. cigarettes)
 - ...
- Laboratory experiments show that ~60% of all trials display the expected behavior (e.g. both participants confess)
- Real-life observations and experimental results show mixed accordance with theory
 - In many cases, the prediction of the model at least has a correct tendency
 - Players do not generally behave as predicted by the model

The Ultimatum Game

- Another empirically well studied sequential two player game
 - Player A may chose any division of an amount of money between him and player B (e.g. 70% for player A and 30% for player B)
 - Player B may ultimately either accept (both receive their share of the money) or reject (neither A nor B receive any money)
- Game theory says its optimal play for A to offer any small amount $\epsilon > 0$ and its optimal play for B to accept any $\epsilon > 0$
 - Assuming the utility of an outcome for each player is proportional to the received amount of money
- However, a large corpus of empirical studies shows that divisions which offer less than 20-30% to player B are rejected
 - Independent of the absolute amount of money at stake
 - Exception: Some tribal communities tend to make very generous offers (and to reject those) to show their prosperity

Some Explanations for Empirical Deviation

- Internal moral standards
 - People behave altruistically, have a sense of fairness → do not offer very unfair splits in Ultimatum
- External moral standards
 - If people feel responsible for their decisions, they make more morally acceptable decisions than if not (e.g. if told that their boss supports non-cooperation in PD, more people chose this option)
- Comparison with others
 - People compare their payoff with the payoff of the other players
 - Social status manifests by actions taken (e.g. submission by accepting a very small offer in Ultimatum)
- We can model this behavior by modifying the utility functions of the players (e.g. do not base utility on the absolute payoff of a game but also on the difference to the other player's payoff)

Repeated Play of PD

- In the real world, we do not play isolated games → do not burn bridges or lose credibility
- Modeled for example in repeated PD (RPD):
 - Play (finite or infinite) sequence of games of PD
 - Utility is sum of sub-game utilities
 - Players observe the outcome of earlier games and can react to it
- In RPD is possible to create credible threats to enforce cooperation
 - “Trigger strategy”: Player A threatens player B to not cooperate in later rounds if B does not cooperate in the current round
 - Is only credible if the potential loss for A in the current round is bigger than the loss in later rounds if the threat is realized
- Contest of computer programs playing repeated PD → Winner strategy “Tit for tat” (In first round, cooperate. In later rounds, do the same as the opponent did previously)

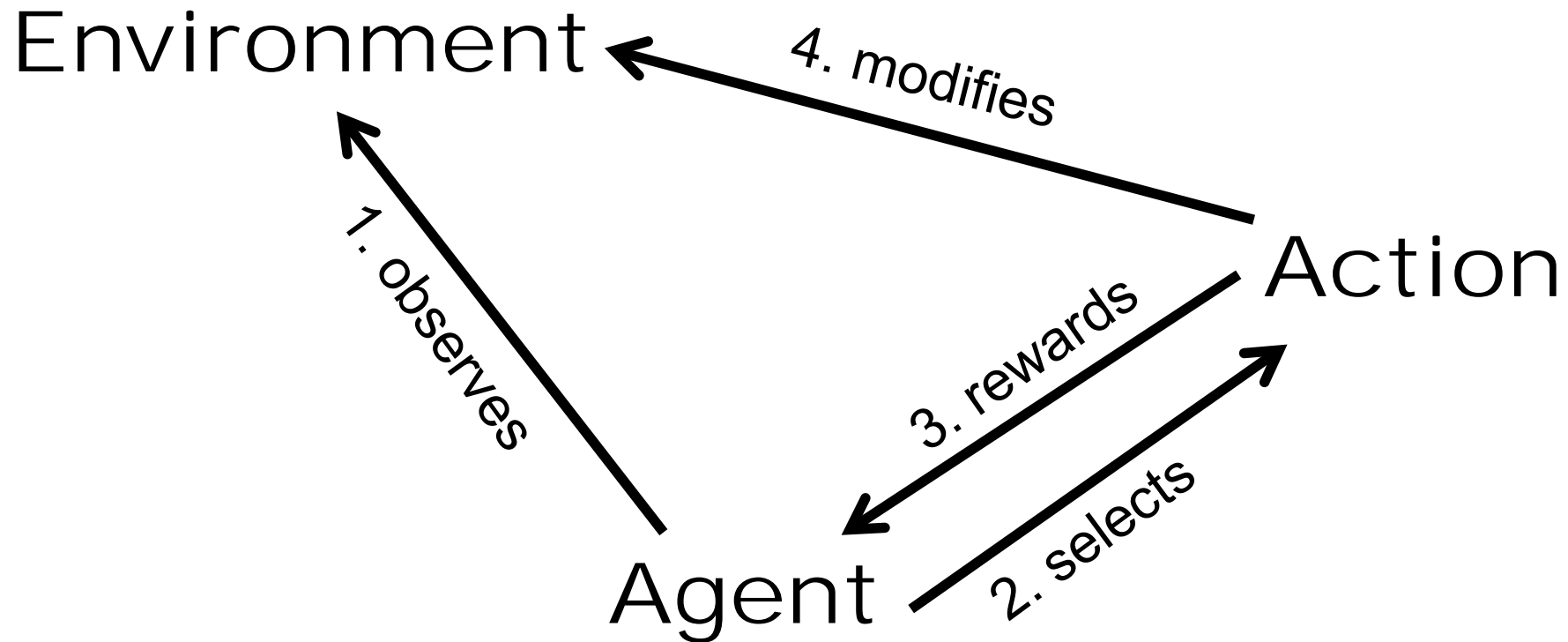
Human Learning

- Classic game theory assumes that each player knows at least his own utility function to make decisions
 - In complex (real-life) scenarios, this may not always be the case
 - We cannot assume that an agent has a-priori knowledge on the result of each action
- Human intelligence manifests in the ability to learn
- Learning is also one of the criteria for a mature cognitive architecture
- Here: strategic learning, planning
- Why study human learning?
 - As model for machine learning
 - To better understand human behavior
- In the following, we concentrate on Reinforcement Learning (RL) as a model of strategic learning

Properties of Human Strategic Learning

- Law of Effect (Thorndike, 1898): Choices that have led to good outcome in the past are more likely to be repeated in the future
 - Implies: Choice behavior is probabilistic
- Power Law of Practice (Blackburn, 1936): Learning curves tend to be steep initially, and then flatter
- Generalization (e.g. Skinner, 1953): Experience of success of a choice is spread to similar choices
- Recency (e.g. Watson, 1930): Recent experience plays a larger role than past experience in determining behavior

Model of Strategic Learning



Markov Decision Process

- A Markov Decision Process (MDP) defines the environment the agent lives in. Consists of:
 - State space S , action space A (here, both finite)
 - Transition function $\delta: S \times A \rightarrow S$ (here, deterministic)
 - Reward function $r: S \times A \rightarrow R$, immediate reward or punishment for each action
- Markovian, i.e. rewards and state transitions depend only on the current state
- For real world scenarios, non-deterministic variants of δ and r are common (yielding probability distributions instead of fixed values)
- An agent in state s selects an action a , observes the resulting new state s' and a corresponding reward $r(s,a)$. This process is repeated until the episode is finished.
 - Optimize reward over a longer course of actions

Optimality Criterion

- Finite horizon model: Optimize the cumulative reward for the state-action sequence of the next N turns:

$$V = \sum_{t=0}^N r(s_t, a_t)$$

- Infinite horizon model: Take into account a potentially infinite sequence of future rewards
- Later rewards have a smaller impact on the planning (higher uncertainty, lesser flexibility):

$$V = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

- Average reward:

$$V = \lim_{h \rightarrow \infty} \frac{1}{h} \sum_{t=0}^h r(s_t, a_t)$$

Evaluating a Strategy

- A (deterministic) strategy is a mapping $\pi: S \rightarrow A$
- A probabilistic strategy defines a probability distribution over all actions for every state
- In the deterministic case, we define

$$V^\pi = \sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)), \text{ where } s_{t+1} = \delta(s_t, \pi(s_t))$$

- If the strategy or the environment are probabilistic, we usually take the expected value
- Define Q-score on level of state-action pairs:

$$Q^\pi(s, a) = r(s, a) + \gamma V^\pi(\delta(s, a))$$

- Q-score measures cumulative reward which is achieved by taking action a in state s (and following π afterwards)
- Optimal strategy: Maximize V for every initial state

RL: A first Algorithm (Value Iteration)

```
Arbitrarily initialize  $V(s)$  for all  $s \in S$   
While strategy not optimal:  
  For each state  $s$ :  
    For each action  $a$ :  
       $Q(s,a) := r(s,a) + \gamma * V(\delta(s,a))$   
     $V(s) := \max_a Q(s,a)$ 
```

- Learned strategy: In state s , select action a , which maximizes the score $V(\delta(s,a))$
- Need to know reward function and transition function \rightarrow Often unrealistic assumption in real-world tasks
- Can we find methods for learning a strategy without knowing the parameters of the underlying MDP? (*model free*)

Q-Learning

- Using some „appropriate“ strategy, the agent explores the state-action space (simple example: random walk)
- It learns from observation of reward and state transitions
- Instead of building an explicit model, it directly learns optimal behavior

```
Arbitrarily initialize  $Q(s,a)$  for all  $s \in S, a \in A$ 
```

```
While strategy not optimal:
```

```
  For current state  $s$ :
```

```
    Pick action  $a$ , leading to state  $s'$  and reward  $r$ 
```

```
     $Q(s,a) := r + \gamma * \max_{a'} Q(s',a')$ 
```

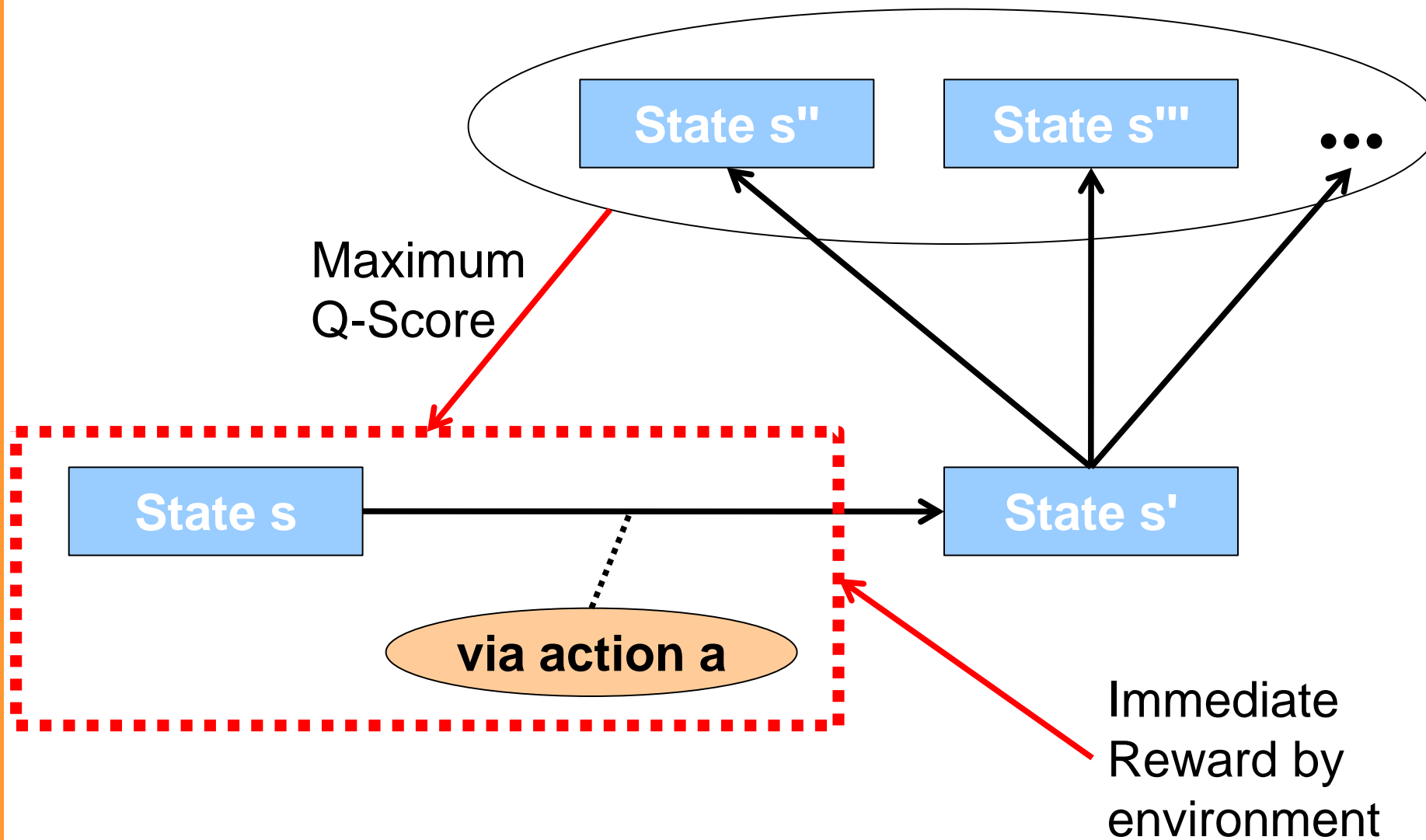
```
     $s := s'$ 
```

- Model free, i.e. does not need a complete representation of reward and transition function

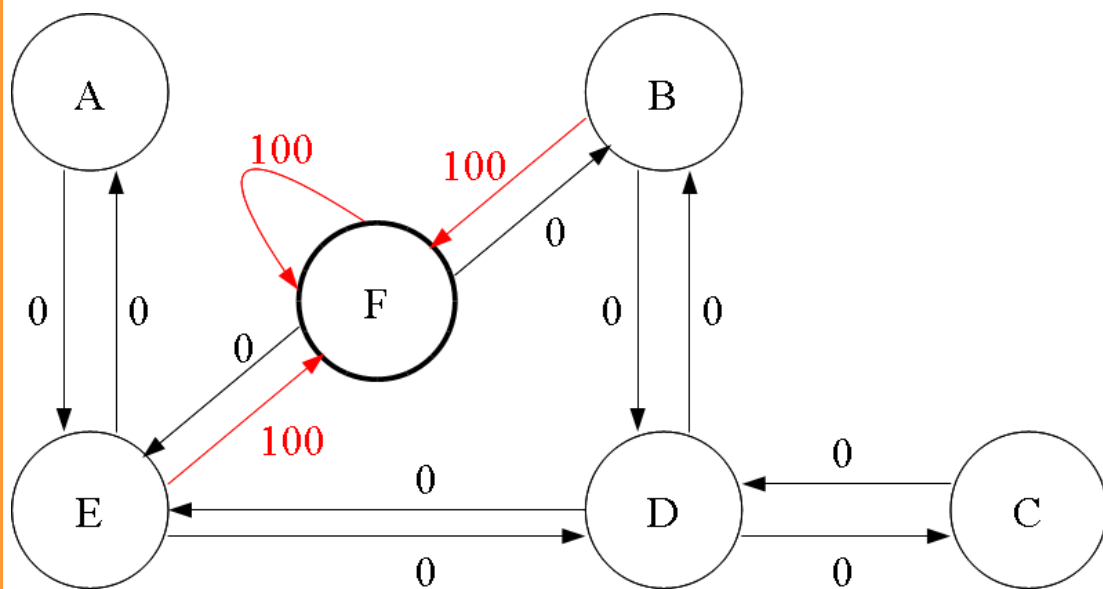
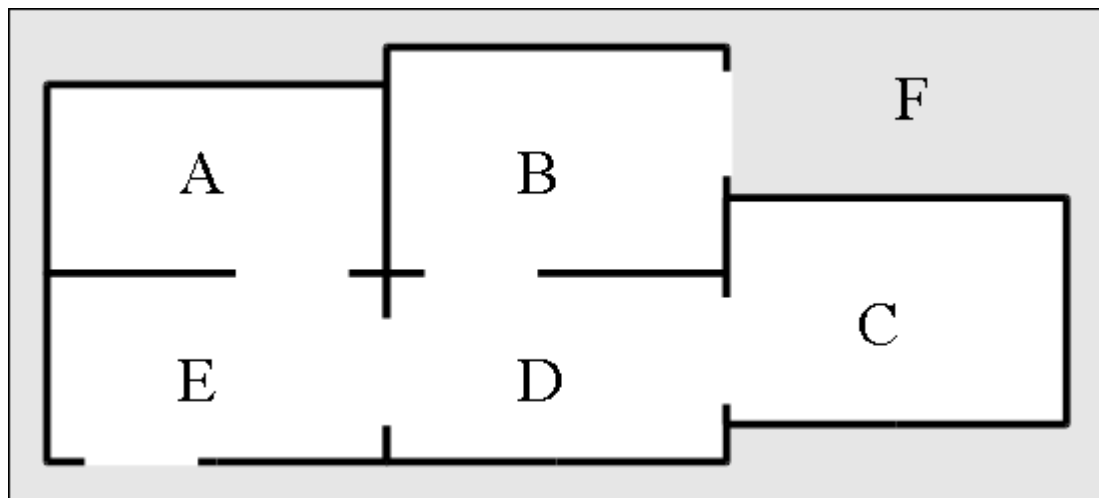
Strategy learned by Q-Learning

- It can be shown that Q-Learning converges to a local maximum when visiting all state-action pairs “often enough”
- When converged, Q-Learning yields an optimal strategy:
 - In a given state s , chose action a which maximizes $Q(s,a)$ (e.g. the expected cumulative reward)
- Q-Learning is *off-policy*: Applied strategy during learning does not need to be the learned strategy
 - Even if during learning the agent walks randomly across the search space, the resulting strategy is optimal

Q-Learning in a Nutshell



Q-Learning: Example

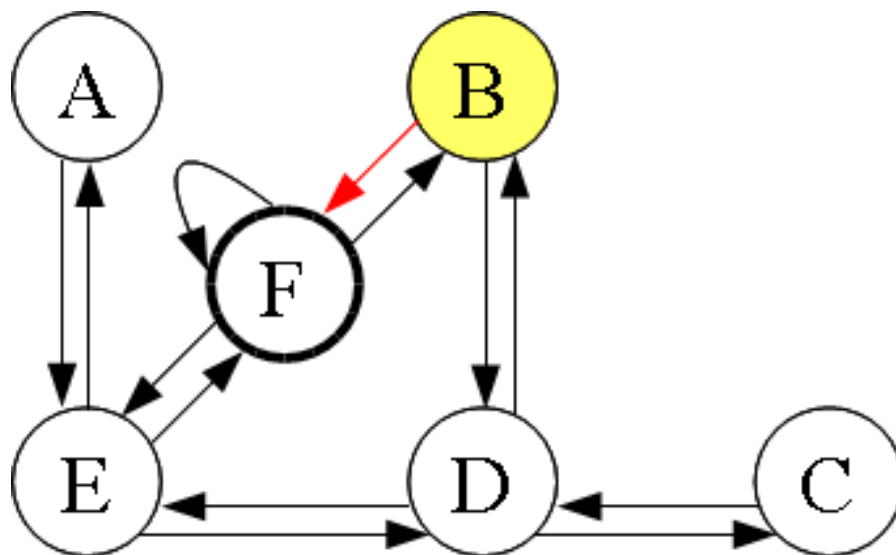


	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	0
C	-	-	-	0	-	-
D	-	0	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q

Q-Learning: Example

1st Episode



	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	0
C	-	-	-	0	-	-
D	-	0	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q



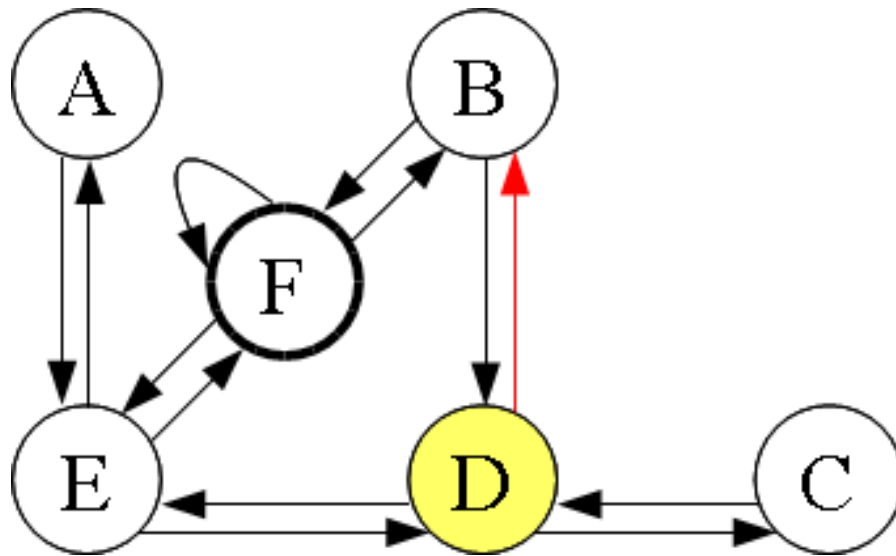
($\gamma=0.8$):

	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	100
C	-	-	-	0	-	-
D	-	0	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q

Q-Learning: Example

2nd Episode (a):



	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	100
C	-	-	-	0	-	-
D	-	0	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q



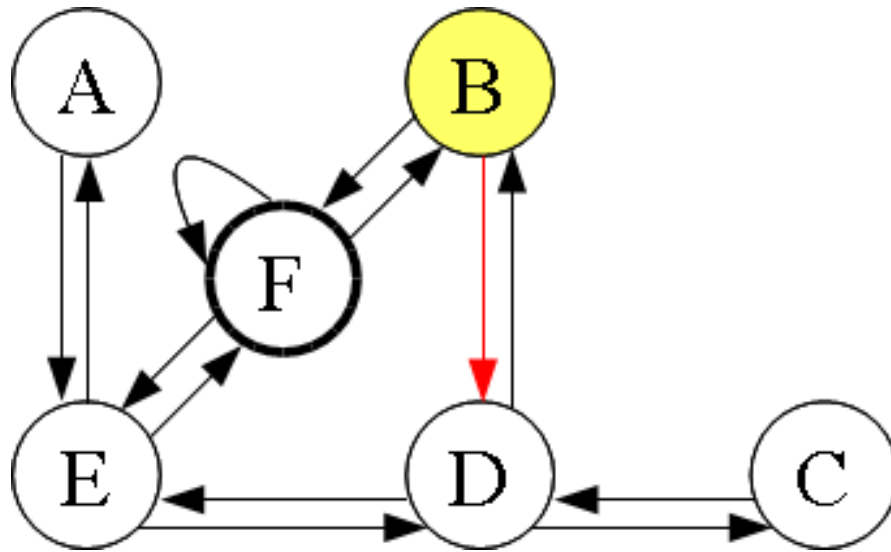
($\gamma=0.8$):

	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	100
C	-	-	-	0	-	-
D	-	80	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q

Q-Learning: Example

2nd Episode (b):



	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	100
C	-	-	-	0	-	-
D	-	80	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q



($\gamma=0.8$):

	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	64	-	100
C	-	-	-	0	-	-
D	-	80	0	-	0	-
E	0	-	-	0	-	0
F	-	0	-	-	0	0

Matrix Q

Q-Learning and Non-Determinism

- Environment is often non-deterministic:
 - Non-deterministic state transition
 - Non-deterministic reward payoff
 - May be caused by
 - Intrinsically non-deterministic environments
 - incomplete representation of reality in state-action-space
- Want to maximize expected cumulative reward
- Solution: Instead of a hard update, just adjust the Q-value towards the new value with *learning factor* α
- $Q(s, a) := r + \gamma * \max_{a'} Q(s', a')$
becomes
$$Q(s, a) := (1 - \alpha) * Q(s, a) + \alpha * (r + \gamma * \max_{a'} Q(s', a'))$$
- α decreases with the number of visits to the updated state-action-pair

Exploration vs. Exploitation

- Need action selection strategy during learning process
 - Random walk will take very long to converge to a useful strategy
- *Exploration* of unknown paths in the state-action space
 - Q-Learning guarantees convergence to local maximum only, if every state-action pair is visited infinitely often
- *Exploitation* of paths in the state-action space known as good
 - Exploitation allows pruning of hopeless parts → faster convergence
- → Need reasonable trade-off between both requirements

Exploration Strategies

- Strategy „ ϵ -greedy“: With probability...
 - $1-\epsilon$: Select optimal action
 - ϵ : select random action (uniformly distributed)
- Strategy „Softmax“: Choose action according to probability according to Boltzmann equation:

$$p_s(a) = \frac{e^{Q(s,a)/t}}{\sum_{a' \in A} e^{Q(s,a')/t}}$$

- t is temperature parameter which monotonically decreases with rising number of training episodes

Learning Environments

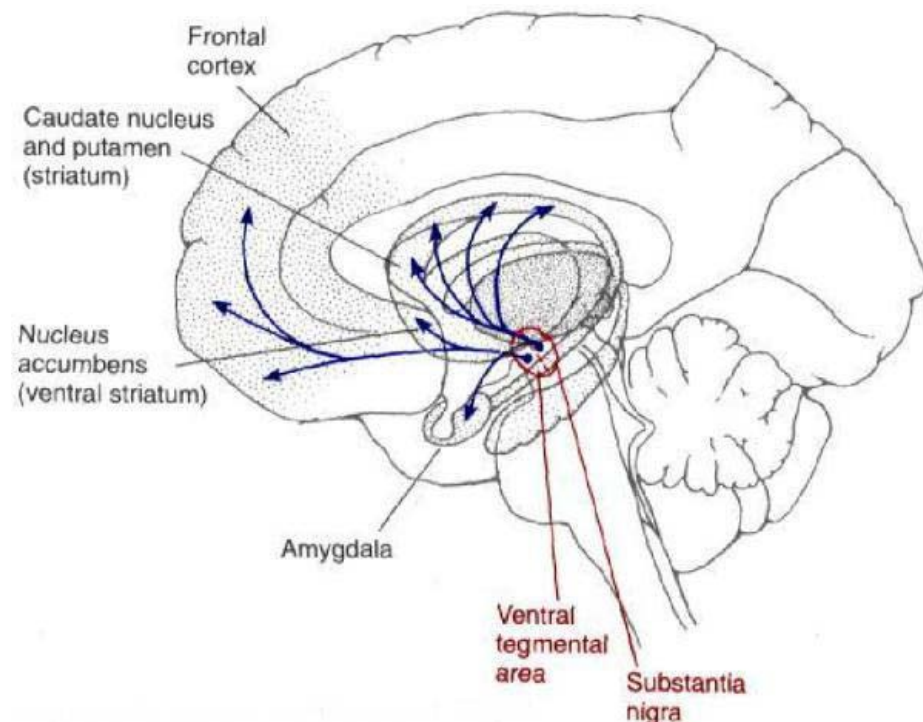
- From where do we derive the rewards and state transitions observed by the learning agent?
- Model:
 - Requires existence of rigorous model of problem domain
 - Only available in simple or very artificial domains
 - Use model-based learning algorithms, no exploration necessary
- Real World:
 - High validity of observations
 - Online learning during lifetime of agent
 - Need acceptable initial strategy if environment is hazardous
- Simulation
 - Compromise between both extremes
 - Simulation is easier to provide than explicit model
 - Allows fast generation of any number of training episodes

Outlook

- Q-Learning is a fundamental RL algorithm, but there is lots more!
 - On-policy learning, e.g. SARSA
 - Model –based approaches and direct policy search, e.g. using gradient descent methods
 - Improvements on learning speed, e.g. eligibility traces or generalization techniques
- For many real-world applications, additional challenges arise
 - Representation of continuous state and action spaces (using function approximators like neural networks)
 - Learning with multiple (adversarial) goals
 - Multi-agent learning
 - Partially observability, e.g. taking uncertainty into account

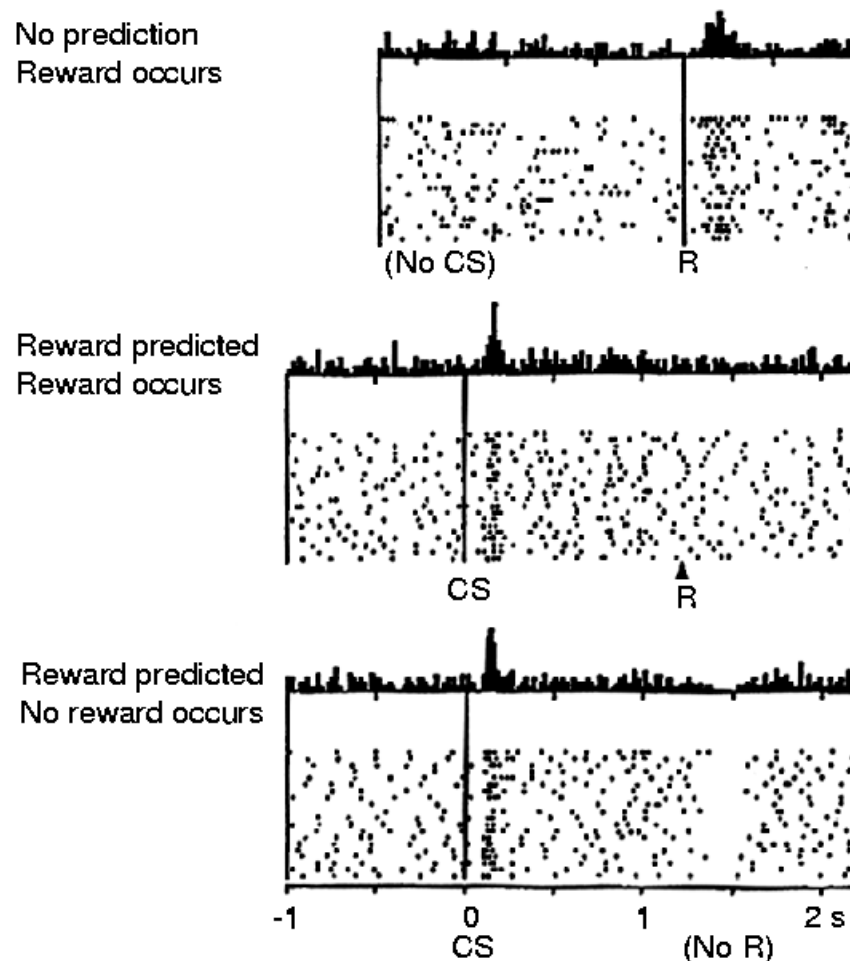
Dopamine

- Dopamine: Neurotransmitter, involved in reward prediction, motor control and attention
- Dopamine-receptive neurons are located in the ventral tegmental area and the substantia nigra
- Connected to striatum, frontal cortex and amygdala



Dopaminic Response to Stimuli and Reward

- Monkeys observed in repeated trials a visual stimulus (CS) and received a delayed reward in form of apple juice (R)



Models of Dopamine Activity

- Dopaminic activity is propagated in time:
 - Before learning, it occurs when reward is received
 - After the learning phase, it occurs when the stimulus is presented
 - In return, activity diminishes for expected reward after the stimulus
 - Inhibition of activity if expected positive reward is missed
- Magnitude of dopaminic response correlates with magnitude of reward
- → It acts as an error signal to reward prediction
$$\Delta(t) = (r(t) + \gamma \hat{V}(t+1)) - \hat{V}(t)$$
- Error signal in RL terms is the difference between the learned value $V(t)$ of the current time step and the observed value
- Learning process can be modeled as Temporal Difference Learning, a type of RL

Observations and Open Problems

- Dopamine also reacts to unexpected neutral events
 - Alternative explanation: Dopamine controls attention on unexpected events regardless of their payoff
 - This does however not explain
- Dopamine does not react to negative rewards (pain, ...)
- Possible counterpart of Dopamine for negative rewards is the Serotonine system
 - Not understood as well as Dopamine

RL and Human Behavior

- Experiment with human participants (Wai-Fat & Anderson, 2006):
 - Humans explore a maze of unknown layout
 - They have to learn correct matching of stimuli and choice options to select the way to the exit
 - Finding the exit is rewarded, Encountering a dead end (after several steps) is penalized
- RL model reflects human learning process
 - Average reward increases over time
 - Choice options at later stages (closer to the rewards) are learned faster than choice options near the entry → reward is propagated backwards like in RL

RL in Cognitive Architectures

- Procedural unit in a cognitive actions selects actions to fulfill the active goal
- Decision criteria in ACT-R:
 - Symbolic: Check preconditions of rules
 - Sub-symbolic: Evaluate utility of each action
- Utility can be learned using RL
 - For utility in ACT-R, we learn the probability of an action to lead to success and its costs (e.g. time to execute) from observation

Discussion

- In this lecture, we discussed two models of human rational decision making
 - Game theory
 - Reinforcement Learning
- Both models are successful in explaining a large body of empirical results and can be extended to explain a large range of phenomena
- However, there are limitations which are hard to circumvent:
 - Heuristics
 - Generic rules which drastically simplify decision making May lead to non-optimal, i.e. non-rational, outcomes
 - Emotion
 - Play an important role in decision making
 - “Gut feeling” is often faster than rational decision making
 - Motivation, Context bias, ...