Multi-Agent Reinforcement Learning

Christian Kalla

Knowledge-Based Systems Group
RWTH Aachen University

8.6.2009/Seminar- Foundations of AI

Outline

The Idea of Reinforcement Learning

- learning by interacting with the environment
- taking actions and receiving rewards
- trying to maximize long-term reward
- mathematical description: Markov Decision Process

Comparison to other Machine Learning Approaches

- **Supervised Learning**
  - given correct input-output pairs
  - correct classification of data given $\rightarrow$ "teacher"
  - example: digit recognition
- **Unsupervised Learning**
  - just "raw data" without labeling given
  - no "teacher"
  - example: clustering methods (k-means,...)
- **Reinforcement Learning**
  - learning by interacting with the environment
  - "natural" approach (related to human learning)
  - feedback from environment in terms of rewards
  - goal: maximize long-term reward

Comparison to other Machine Learning Approaches

- **Supervised Learning**
  - given correct input-output pairs
  - correct classification of data given → "teacher"
  - example: digit recognition
- **Unsupervised Learning**
  - just "raw data" without labeling given
  - no "teacher"
  - example: clustering methods (k-means,...)
- **Reinforcement Learning**
  - learning by interacting with the environment
  - "natural" approach (related to human learning)
  - feedback from environment in terms of rewards
  - goal: maximize long-term reward

Comparison to other Machine Learning Approaches

- **Supervised Learning**
  - given correct input-output pairs
  - correct classification of data given → "teacher"
  - example: digit recognition
- **Unsupervised Learning**
  - just "raw data" without labeling given
  - no "teacher"
  - example: clustering methods (k-means,...)
- **Reinforcement Learning**
  - learning by interacting with the environment
  - "natural" approach (related to human learning)
  - feedback from environment in terms of rewards
  - goal: maximize long-term reward

Comparison to other Machine Learning Approaches

- **Supervised Learning**
  - given correct input-output pairs
  - correct classification of data given $\rightarrow$ "teacher"
  - example: digit recognition
- **Unsupervised Learning**
  - just "raw data" without labeling given
  - no "teacher"
  - example: clustering methods (k-means,...)
- **Reinforcement Learning**
  - learning by interacting with the environment
  - "natural" approach (related to human learning)
  - feedback from environment in terms of rewards
  - goal: maximize long-term reward

Outline

The Markov Property

Markov Property

$$Pr\left\{s_{t+1}=s', r_{t+1}=r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, ..., r_1, s_0, a_0\right\}$$
$$=Pr\left\{s_{t+1}=s', r_{t+1}=r | s_t, a_t\right\}$$

The probability distribution of the next state only depends on the previous state and not on all the states visited before

Markov Decision Process(MDP)

Components of an MDP

- a set of states $\mathcal{S}$
- a set of actions $\mathcal{A}$
- a set of rewards $\mathfrak{R}$
- a transition function $T : \mathcal{S} \times \mathcal{A} \rightarrow PD(\mathcal{S})$ where $PD(\mathcal{S})$ denotes the set of probability distribution over $\mathcal{S}$
- a reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathfrak{R}$

Goal:Maximize expected sum of discounted future rewards:

$$E\left\{\sum_{j=0}^{\infty} \gamma^j r_{t+j}\right\} \tag{1}$$

Categorization of Algorithms

The Q-Learning Algorithm

Q-Learning Algorithm

Initialize $Q(s, a)$ arbitrarily
for each episode **do**
  Initialize s
  repeat
    Choose a from s using policy derived from Q
    Take action a and observe a, s'
    $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$
    $s \leftarrow s'$
  until s is terminal
end for

Yet Another Slide

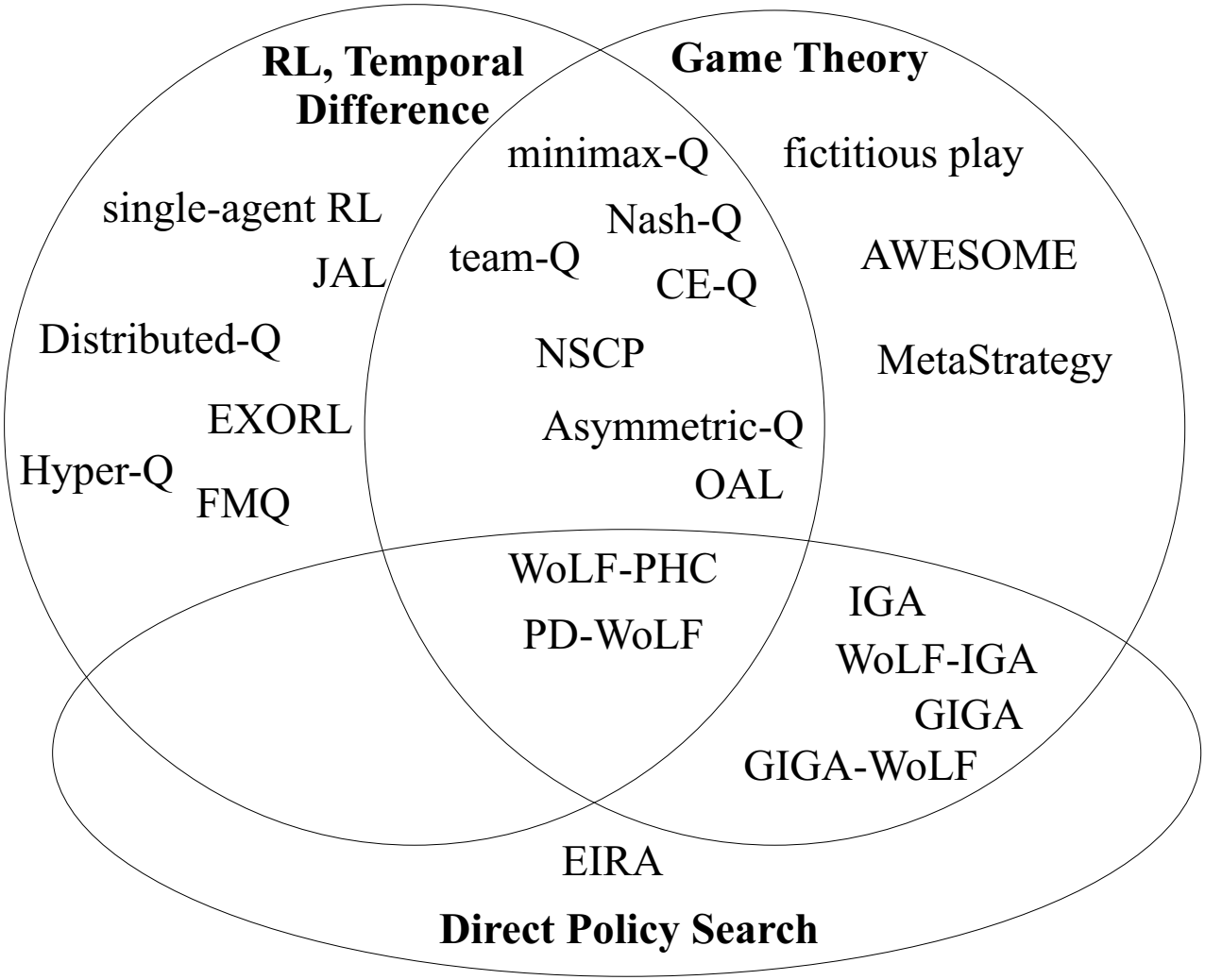
1st column

- Item1
- Item2
- ...

2nd column

Outline

Overview of MARL algorithms



A Robot Soccer Scenario

The Minimax-Q Learning Algorithm

Last Slide

▸ Q Learning applet