CS 747, Autumn 2022: Lecture 12

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

Reinforcement Learning

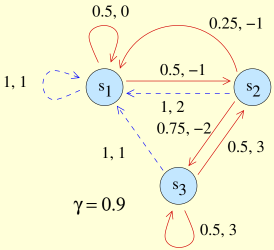
1. Reinforcement learning problem
2. Upcoming topics
3. Applications

Reinforcement Learning

1. Reinforcement learning problem
2. Upcoming topics
3. Applications

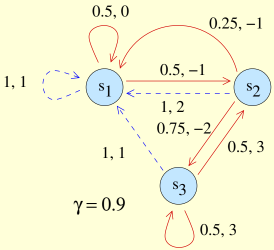
The Learning Setting

Underlying MDP:

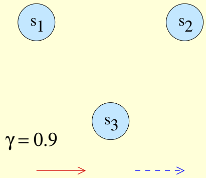


The Learning Setting

Underlying MDP:

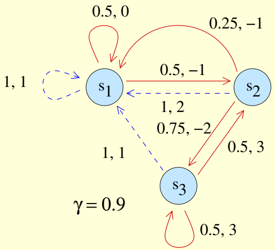


Agent's view:

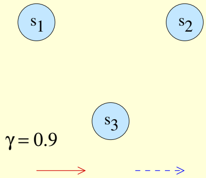


The Learning Setting

Underlying MDP:



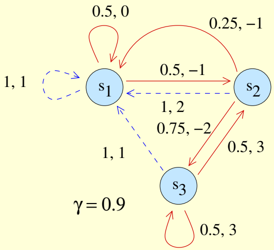
Agent's view:



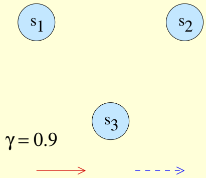
- From current state, agent takes action.

The Learning Setting

Underlying MDP:



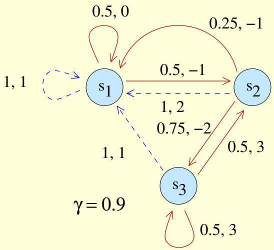
Agent's view:



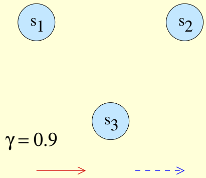
- From current state, agent takes action.
- Environment (MDP) decides next state and reward.

The Learning Setting

Underlying MDP:



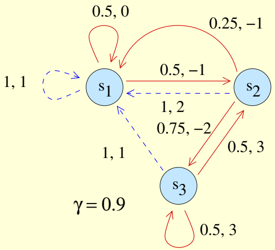
Agent's view:



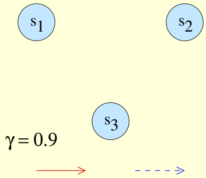
- From current state, agent takes action.
- Environment (MDP) decides next state and reward.
- Possible history: s_2, RED, -2, s_3, BLUE, 1, s_1, RED, 0, s_1, ...

The Learning Setting

Underlying MDP:



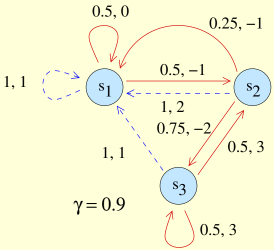
Agent's view:



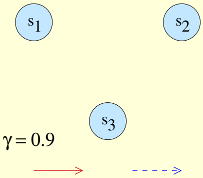
- From current state, agent takes action.
- Environment (MDP) decides next state and reward.
- Possible history: s_2, RED, -2, s_3, BLUE, 1, s_1, RED, 0, s_1, ...
- History conveys information about the MDP to the agent.

The Learning Setting

Underlying MDP:



Agent's view:



- From current state, agent takes action.
- Environment (MDP) decides next state and reward.
- Possible history: s_2, RED, -2, s_3, BLUE, 1, s_1, RED, 0, s_1, ...
- History conveys information about the MDP to the agent.
  Can the agent eventually take optimal actions?

Planning and Learning

- In the planning setting, the entire MDP (S, A, T, R, γ) is available as an input. Obtaining $\pi^\star$ is a computational problem.

Planning and Learning

- In the planning setting, the entire MDP (S, A, T, R, γ) is available as an input. Obtaining $\pi^\star$ is a computational problem.

- In the learning setting, the agent only knows S, A, γ, and sometimes R. It has to make inferences about T (and sometimes R) by taking actions from different states.

Planning and Learning

- In the planning setting, the entire MDP (S, A, T, R, γ) is available as an input. Obtaining $\pi^\star$ is a computational problem.

- In the learning setting, the agent only knows S, A, γ, and sometimes R. It has to make inferences about T (and sometimes R) by taking actions from different states.

- For $t \geq 0$, let $h^t = (s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t)$ denote a t-length history.

Planning and Learning

- In the planning setting, the entire MDP (S, A, T, R, γ) is available as an input. Obtaining $\pi^\star$ is a computational problem.

- In the learning setting, the agent only knows S, A, γ, and sometimes R. It has to make inferences about T (and sometimes R) by taking actions from different states.

- For $t \geq 0$, let $h^t = (s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t)$ denote a t-length history.
- A learning algorithm L is a mapping from the set of all histories to the set of all (probability distributions over) arms.

Planning and Learning

- In the planning setting, the entire MDP (S, A, T, R, γ) is available as an input. Obtaining $\pi^\star$ is a computational problem.

- In the learning setting, the agent only knows S, A, γ, and sometimes R. It has to make inferences about T (and sometimes R) by taking actions from different states.

- For $t \geq 0$, let $h^t = (s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t)$ denote a t-length history.
- A learning algorithm L is a mapping from the set of all histories to the set of all (probability distributions over) arms.
- **Learning problem**: Can we construct L such that

$$\lim_{H \to \infty} \frac{1}{H} \left(\sum_{t=0}^{H-1} \mathbb{P}\{a^t \sim L(h^t) \text{ is an optimal action for } s^t\} \right) = 1?$$

Reinforcement Learning

1. Reinforcement learning problem
2. Upcoming topics
3. Applications

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.
- Model-based RL
  - Using (approximate) representations of T and R for learning.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.
- Model-based RL
  - Using (approximate) representations of T and R for learning.
- Batch RL
  - Storing and learning from a sequence of transitions (batch).

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.
- Model-based RL
  - Using (approximate) representations of T and R for learning.
- Batch RL
  - Storing and learning from a sequence of transitions (batch).
- Monte Carlo tree search
  - Planning for action selection.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.
- Model-based RL
  - Using (approximate) representations of T and R for learning.
- Batch RL
  - Storing and learning from a sequence of transitions (batch).
- Monte Carlo tree search
  - Planning for action selection.
- Multiagent RL
  - Coping with other learning agents.

Upcoming Topics

- Temporal difference learning: prediction and control
  - On-line estimation of value function/action value function.
- Generalisation and function approximation
  - Compact representations to handle large state spaces.
- Policy gradient and policy search methods
  - Direct search over policy parameters.
- Model-based RL
  - Using (approximate) representations of T and R for learning.
- Batch RL
  - Storing and learning from a sequence of transitions (batch).
- Monte Carlo tree search
  - Planning for action selection.
- Multiagent RL
  - Coping with other learning agents.
- Applications
  - ATARI games (Mnih *et al.* (2015)), Go (Silver *et al.* (2016)).

Reinforcement Learning

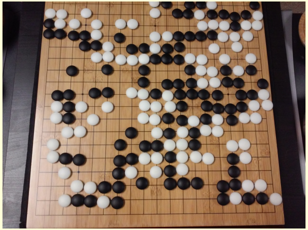
1. Reinforcement learning problem
2. Upcoming topics
3. Applications

Board Games

Backgammon

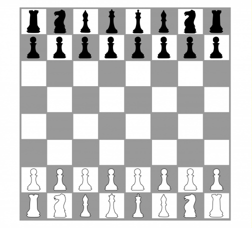
[1]

Go



[2]

Chess



[3]

References: Tesauro (1992), Silver *et al.* (2018).

1. https://www.publicdomainpictures.net/pictures/60000/velka/backgammon.jpg.
2. https://www.publicdomainpictures.net/pictures/170000/velka/finished-go-game.jpg.
3. https://www.publicdomainpictures.net/pictures/80000/velka/chess-board-and-pieces.jpg.

Robotics and Control

[1]

Reference: Ng *et al.* (2003).

1. `https://www.publicdomainpictures.net/pictures/20000/velka/police-helicopter-8712919948643Mk.jpg`.

Video Games

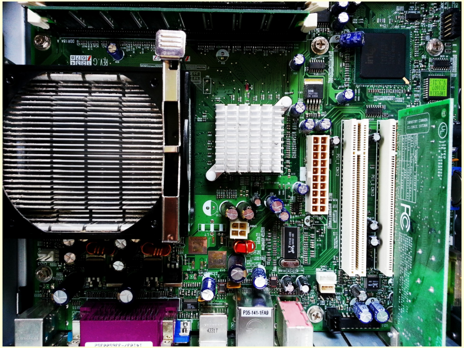
[1]

Reference: Mnih *et al.* (2015).

1. `https://www.publicdomainpictures.net/pictures/30000/velka/arcade-gaming.jpg`.

Computer Systems

Optimising a memory controller



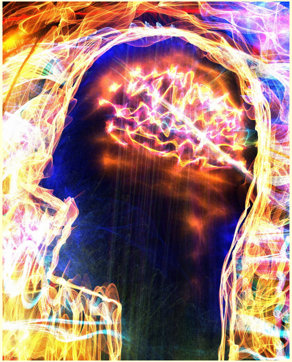
[1]

- Reference: İpek *et al.* (2008).

1. https://www.publicdomainpictures.net/pictures/100000/velka/motherboard.jpg.

Healthcare

Adaptive treatment of epilepsy



[1]

- Reference: Guez *et al.* (2008).

1. https://www.publicdomainpictures.net/pictures/140000/velka/brain-signals.jpg.

Finance

Stock trading

- Reference: Moody and Saffell (2001).

Reinforcement Learning

1. Reinforcement learning problem

2. Upcoming topics

3. Applications