CS394R
Reinforcement Learning: Theory and Practice

Peter Stone

Department of Computer Science
The University of Texas at Austin
BE a reinforcement learner
BE a reinforcement learner

- You, as a class, act as a learning agent
BE a reinforcement learner

- You, as a class, act as a learning agent
- **Actions**: Wave, Stand, Clap
BE a reinforcement learner

• You, as a class, act as a learning agent

• **Actions**: Wave, Stand, Clap

• **Observations**: colors, reward
BE a reinforcement learner

- You, as a class, act as a learning agent

- **Actions**: Wave, Stand, Clap

- **Observations**: colors, reward

- **Goal**: Find an optimal *policy*
BE a reinforcement learner

- You, as a class, act as a learning agent
- **Actions**: Wave, Stand, Clap
- **Observations**: colors, reward
- **Goal**: Find an optimal *policy*
  - Way of selecting actions that gets you the most reward
How did you do it?
How did you do it?

- What is your policy?
- What does the world look like?
Formalizing What Just Happened

Knowns:
Formalizing What Just Happened

Knouns:

- $\mathcal{O} = \{\text{Blue, Red, Green, Black, ...}\}
- \text{Rewards in } \mathbb{R}
- \mathcal{A} = \{\text{Wave, Clap, Stand}\}$
Formalizing What Just Happened

Knowns:

- $O = \{\text{Blue, Red, Green, Black, …}\}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$

\[ o_0, a_0, r_0, o_1, a_1, r_1, o_2, \ldots \]
Formalizing What Just Happened

Knowns:

- $\mathcal{O} = \{\text{Blue}, \text{Red}, \text{Green}, \text{Black}, \ldots\}$
- Rewards in $\mathbb{R}$
- $\mathcal{A} = \{\text{Wave, Clap, Stand}\}$

\[ o_0, a_0, r_0, o_1, a_1, r_1, o_2, \ldots \]

Unknowns:
Formalizing What Just Happened

Knowns:
- $O = \{\text{Blue, Red, Green, Black, \ldots}\}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$
- $o_0, a_0, r_0, o_1, a_1, r_1, o_2, \ldots$

Unknowns:
- $S = 4 \times 3$ grid
- $R : S \times A \mapsto \mathbb{R}$
- $T : S \mapsto O$
- $P : S \times A \mapsto S$
Formalizing What Just Happened

**Knowns:**
- \( \mathcal{O} = \{ \text{Blue, Red, Green, Black, \ldots} \} \)
- Rewards in \( \mathbb{R} \)
- \( \mathcal{A} = \{ \text{Wave, Clap, Stand} \} \)

**Unknowns:**
- \( \mathcal{S} = 4 \times 3 \) grid
- \( \mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \)
- \( \mathcal{T} = \mathcal{S} \rightarrow \mathcal{O} \)
- \( \mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \)

\[ s_0, o_0, a_0, r_0, s_1, o_1, a_1, r_1, s_2, o_2, \ldots \]
Formalizing What Just Happened

Knowns:
- \( \mathcal{O} = \{ \text{Blue}, \text{Red}, \text{Green}, \text{Black}, \ldots \} \)
- Rewards in \( \mathbb{R} \)
- \( \mathcal{A} = \{ \text{Wave, Clap, Stand} \} \)

Unknowns:
- \( \mathcal{S} = 4 \times 3 \) grid
- \( \mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \)
- \( \mathcal{T} = \mathcal{S} \rightarrow \mathcal{O} \)
- \( \mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \)

\( s_0, o_0, a_0, r_0, s_1, o_1, a_1, r_1, s_2, o_2, \ldots \)

\( o_i = \mathcal{T}(s_i) \)
Formalizing What Just Happened

**Knowns:**
- \( \mathcal{O} = \{ \text{Blue, Red, Green, Black, \ldots} \} \)
- Rewards in \( \mathbb{R} \)
- \( \mathcal{A} = \{ \text{Wave, Clap, Stand} \} \)
  
  \[
  o_0, a_0, r_0, o_1, a_1, r_1, o_2, \ldots
  \]

**Unknowns:**
- \( \mathcal{S} = 4 \times 3 \) grid
- \( \mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \)
- \( \mathcal{T} = \mathcal{S} \rightarrow \mathcal{O} \)
- \( \mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \)
  
  \[
  s_0, o_0, a_0, r_0, s_1, o_1, a_1, r_1, s_2, o_2, \ldots
  \]

\[
\begin{align*}
o_i & = \mathcal{T}(s_i) \\
r_i & = \mathcal{R}(s_i, a_i)
\end{align*}
\]
Formalizing What Just Happened

Knowns:

- $\mathcal{O} = \{\text{Blue, Red, Green, Black, …}\}$
- Rewards in $\mathbb{R}$
- $\mathcal{A} = \{\text{Wave, Clap, Stand}\}$

Unknows:

- $\mathcal{S} = 4\times3$ grid
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$
- $\mathcal{T} : \mathcal{S} \mapsto \mathcal{O}$
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$
This Course

- Reinforcement Learning theory (start)
This Course

- Reinforcement Learning theory (start)
- Reinforcement Learning in practice (end)
The Big Picture

- AI
The Big Picture

- AI $\rightarrow$ ML
The Big Picture

• AI $\rightarrow$ ML $\rightarrow$ RL
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

- Types of Machine Learning
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

- Types of Machine Learning
  
  Supervised learning: learn from labeled examples
The Big Picture

• AI $\rightarrow$ ML $\rightarrow$ RL

• Types of Machine Learning

  Supervised learning: learn from labeled examples
  Unsupervised learning: cluster unlabeled examples
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

- Types of Machine Learning

  **Supervised learning:** learn from labeled examples
  **Unsupervised learning:** cluster unlabeled examples
  **Reinforcement learning:** learn from interaction
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

- Types of Machine Learning
  
  **Supervised learning:** learn from labeled examples
  
  **Unsupervised learning:** cluster unlabeled examples
  
  **Reinforcement learning:** learn from interaction
    
    - Defined by the problem
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

- Types of Machine Learning

  **Supervised learning:** learn from labeled examples
  **Unsupervised learning:** cluster unlabeled examples
  **Reinforcement learning:** learn from interaction
    - Defined by the problem
    - Many approaches possible (including evolutionary)
The Big Picture

- AI $\rightarrow$ ML $\rightarrow$ RL

Types of Machine Learning

** Supervised learning:** learn from labeled examples
** Unsupervised learning:** cluster unlabeled examples
** Reinforcement learning:** learn from interaction
  - Defined by the problem
  - Many approaches possible (including evolutionary)
  - Book focuses on a particular class of approaches
Reduced Formalism

Knowns:

- $S = \{\text{Blue, Red, Green, Black, \ldots}\}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$

\[
\begin{align*}
S_0, a_0, r_0, S_1, a_1, r_1, S_2, \ldots
\end{align*}
\]
Reduced Formalism

**Knowns:**

- $S = \{\text{Blue, Red, Green, Black, } \ldots \}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$

```
S_0, a_0, r_0, S_1, a_1, r_1, S_2, \ldots
```

**Unknowns:**
Reduced Formalism

Knowns:
- $S = \{\text{Blue, Red, Green, Black, \ldots}\}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$

Unknwns:
- $R : S \times A \rightarrow \mathbb{R}$
- $P : S \times A \rightarrow S$
Reduced Formalism

Knowns:

- $S = \{\text{Blue, Red, Green, Black,} \ldots\}$
- Rewards in $\mathbb{R}$
- $A = \{\text{Wave, Clap, Stand}\}$

Unknows:

- $R : S \times A \mapsto \mathbb{R}$
- $P : S \times A \mapsto S$

\[ r_i = R(s_i, a_i) \quad s_{i+1} = P(s_i, a_i) \]
This course

- Agent’s perspective: only \textit{policy} under control
  - State representation, reward function given
  - Focus on policy algorithms, theoretical analyses
This course

- Agent’s perspective: only policy under control
  - State representation, reward function given
  - Focus on policy algorithms, theoretical analyses
  - Appeal: program by just specifying goals
This course

- Agent’s perspective: only **policy** under control
  - State representation, reward function given
  - Focus on policy algorithms, theoretical analyses
  - Appeal: program by just specifying goals
  - Practice: need to pick the representation, reward
This course

- Agent’s perspective: only policy under control
  - State representation, reward function given
  - Focus on policy algorithms, theoretical analyses
  - Appeal: program by just specifying goals
  - Practice: need to pick the representation, reward
  - videos
This course

• Agent’s perspective: only policy under control
  – State representation, reward function given
  – Focus on policy algorithms, theoretical analyses
  – Appeal: program by just specifying goals
  – Practice: need to pick the representation, reward
  – videos

• Methodical approach
  – Solid foundation rather than comprehensive coverage
This course

- Agent’s perspective: only **policy** under control
  - State representation, reward function given
  - Focus on policy algorithms, theoretical analyses
  - Appeal: program by just specifying goals
  - Practice: need to pick the representation, reward
  - videos

- Methodical approach
  - Solid foundation rather than comprehensive coverage
  - RL reading group
Syllabus

- Available on-line
BREAK TIME!
BREAK TIME!

• Bon appetit!
Good Morning Colleagues
Good Morning Colleagues

- Are there any questions?
Logistics
Logistics

- Nice responses!
Logistics

- Nice responses!
  - Length and content good
Logistics

- Nice responses!
  - Length and content good
  - Be clear and specific
Logistics

- Nice responses!
  - Length and content good
  - Be clear and specific
  - Look for programming assignment opportunities
Logistics

- Nice responses!
  - Length and content good
  - Be clear and specific
  - Look for programming assignment opportunities
  - I have author’s responses to exercises
Logistics

• Nice responses!
  – Length and content good
  – Be clear and specific
  – Look for programming assignment opportunities
  – I have author’s responses to exercises

• Programming language
Logistics

- Nice responses!
  - Length and content good
  - Be clear and specific
  - Look for programming assignment opportunities
  - I have author’s responses to exercises

- Programming language

- Self-introductions
Some Questions

- Reward function vs. value function
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
  - Phil making breakfast example
Some Questions

• Reward function vs. value function
  – Tic-tac-toe example
  – Phil making breakfast example

• Could the reward function be learned/altered?
Some Questions

• Reward function vs. value function
  – Tic-tac-toe example
  – Phil making breakfast example

• Could the reward function be learned/altered?

• Tic-tac-toe example: what are the converged values?
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
  - Phil making breakfast example

- Could the reward function be learned/altered?

- Tic-tac-toe example: what are the converged values?

- What happens in self play?
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
  - Phil making breakfast example

- Could the reward function be learned/altered?

- Tic-tac-toe example: what are the converged values?

- What happens in self play?

- How and when to explore?
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
  - Phil making breakfast example

- Could the reward function be learned/altered?

- Tic-tac-toe example: what are the converged values?

- What happens in self play?

- How and when to explore?

- Role of step size
Some Questions

- Reward function vs. value function
  - Tic-tac-toe example
  - Phil making breakfast example
- Could the reward function be learned/altered?
- Tic-tac-toe example: what are the converged values?
- What happens in self play?
- How and when to explore?
- Role of step size
- Does speed of learning matter?
Some Questions

• Distinction with evolutionary methods?
  – Tic-tac-toe example
Some Questions

- Distinction with evolutionary methods?
  - Tic-tac-toe example
  - Phil making breakfast example
Some Questions

- Distinction with evolutionary methods?
  - Tic-tac-toe example
  - Phil making breakfast example

- Is evolutionary learning ever better?
Some Questions

- Distinction with evolutionary methods?
  - Tic-tac-toe example
  - Phil making breakfast example
- Is evolutionary learning ever better?
- Distinguishing features (from supervised learning)?
Some Questions

- Distinction with evolutionary methods?
  - Tic-tac-toe example
  - Phil making breakfast example

- Is evolutionary learning ever better?

- Distinguishing features (from supervised learning)?
  - trial-error search, delayed reward
  - exploration vs. exploitation (chapt. 2)
Assignments

- Join piazza!
Assignments

- Join piazza!
- Read Chapters 2 and 3 (and 1 if you haven’t)
Assignments

• Join piazza!

• Read Chapters 2 and 3 (and 1 if you haven’t)

• Send a reading response by 1pm Tuesday
Assignments

• Join piazza!
• Read Chapters 2 and 3 (and 1 if you haven’t)
• Send a reading response by 1pm Tuesday
• Need a discussion leader volunteer and experiment presenter