Teaching Teammates in Ad Hoc Teams

Peter Stone
Director, Learning Agents Research Group
Department of Computer Sciences
The University of Texas at Austin

Joint work with
Gal A. Kaminka, Sarit Kraus, Bar Ilan University
Jeffrey S. Rosenschein, Hebrew University
Research Question

To what degree can autonomous intelligent agents **learn** in the presence of teammates and/or adversaries in real-time, **dynamic** domains?
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

- Autonomous agents
- Robotics
- Machine learning (RL)
- Multiagent systems
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

- Autonomous agents
- Robotics
- Machine learning (RL)
- Multiagent systems
  - e-commerce
  - mechanism design

© 2009 Peter Stone
Research Question

To what degree can autonomous intelligent agents learn in the presence of teammates and/or adversaries in real-time, dynamic domains?

- Autonomous agents
- Robotics
- Machine learning (RL)
- Multiagent systems
  - e-commerce
  - mechanism design
Teamwork
Teamwork
Teamwork

• Typical scenario: pre-coordination
  – People practice together
  – Robots given coordination languages, protocols
  – “Locker room agreement” (Stone & Veloso, ’99)
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

**Goal:** Create a good team player
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

**Goal:** Create a good team player

- Minimal representative scenarios
  - One teammate, no communication
  - Fixed and known behavior
Scenarios

- Cooperative normal form game (w/ Kaminka & Rosenschein)

\[
\begin{array}{c|ccc}
M1 & b_0 & b_1 & b_2 \\
\hline
a_0 & 25 & 1 & 0 \\
a_1 & 10 & 30 & 10 \\
a_2 & 0 & 33 & 40 \\
\end{array}
\]

- Cooperative $k$-armed bandit (w/ Kraus)
Scenarios

• Cooperative normal form game (w/ Kaminka & Rosenschein)

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

• Cooperative $k$-armed bandit (w/ Kraus)
Formalism

- Agent A in our control: actions $a_0, a_1, \ldots a_{x-1}$
- Agent B reacts in a fixed way: $b_0, b_1, \ldots, b_{y-1}$
Formalism

- Agent A in our control: actions $a_0, a_1, \ldots a_{x-1}$
- Agent B reacts in a fixed way: $b_0, b_1, \ldots, b_{y-1}$
- **Game theory**: normal form, fully cooperative

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Payoff from joint action $(a_i, b_j)$: $m_{i,j}$
Formalism

- Agent A in our control: actions $a_0, a_1, \ldots a_{x-1}$
- Agent B reacts in a fixed way: $b_0, b_1, \ldots, b_{y-1}$
- **Game theory**: normal form, fully cooperative

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Payoff from joint action $(a_i, b_j)$: $m_{i,j}$
- Highest payoff $m^*$ always at $(a_{x-1}, b_{y-1})$
- **Agent B**’s default action: $b_0$
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B not adaptive $\implies a_0$ always

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Objective

- **Agent A**’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on **Agent B**’s strategy

  **Agent B** not adaptive $\implies a_0$ always

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, …
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B best response $\Rightarrow$ can do better

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, ...
- Reward sequence: 25
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B best response $\Rightarrow$ can do better

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, ...
- Reward sequence: 25, 0
Objective

• Agent A’s goal: action sequence with highest reward
  – Undiscounted, medium-term (finite)
  – Depends on Agent B’s strategy

  **Agent B best response \(\Rightarrow\) can do better**

<table>
<thead>
<tr>
<th>(M1)</th>
<th>(b_0)</th>
<th>(b_1)</th>
<th>(b_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_0)</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(a_1)</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>(a_2)</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

• Reward sequence: \(25, 25, 25, \ldots\)
• Reward sequence: \(25, 0, 40\)
Objective

• Agent A’s goal: action sequence with highest reward
  – Undiscounted, medium-term (finite)
  – Depends on Agent B’s strategy

  Agent B best response $\implies$ can do better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

• Reward sequence: 25, 25, 25, ...
• Reward sequence: 25, 0, 40, 40, 40, ...
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

  Agent B best response $\Rightarrow$ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, …
- Reward sequence: 25, 0, 40, 40, 40, …
- How?
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B best response ➞ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, . . .
- Reward sequence: 25, 0, 40, 40, 40, . . .
- Reward sequence: 25, 10
Objective

- **Agent A’s goal:** action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

**Agent B best response** $\Rightarrow$ or even better

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, ...
- Reward sequence: 25, 0, 40, 40, 40, ...
- Reward sequence: 25, 10, 33
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B best response $\Rightarrow$ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, …
- Reward sequence: 25, 0, 40, 40, 40, …
- Reward sequence: 25, 10, 33, 40
Objective

• **Agent A’s goal:** action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on **Agent B’s strategy**

  *Agent B best response* $\implies$ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

• Reward sequence: 25, 25, 25, ...
• Reward sequence: 25, 0, 40, 40, 40, ...
• Reward sequence: 25, 10, 33, 40, 40, 40, ...
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

\[ \text{Agent B best response} \Rightarrow \text{or even better} \]

<table>
<thead>
<tr>
<th>( M1 )</th>
<th>( b_0 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, …
- Reward sequence: 25, 0, 40, 40,… (65 from 1st 3)
- Reward sequence: 25, 10, 33, 40,… (68 from 1st 3)
Objective

• Agent A’s goal: action sequence with highest reward
  – Undiscounted, medium-term (finite)
  – Depends on Agent B’s strategy

Agent B best response $\implies$ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

• Reward sequence: 25, 25, 25, . . .
• Reward sequence: 25, 0, 40, 40, . . .
• Reward sequence: 25, 10, 33, 40, . . .

Cost: $15+40=55$
Cost: $15+30+7=52$
Objective

- Agent A’s goal: action sequence with highest reward
  - Undiscounted, medium-term (finite)
  - Depends on Agent B’s strategy

Agent B best response $\Rightarrow$ or even better

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Reward sequence: 25, 25, 25, . . .
- Reward sequence: 25, 0, 40, 40, . . . Cost: 55, Length: 2
- Reward sequence: 25, 10, 33, 40, . . . Cost: 52, Length: 3
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy

   - \textit{mem}: memory size
   - $\epsilon$: degree of randomness
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy
   - $\text{mem}$: memory size
   - $\epsilon$: degree of randomness

2. Agent A knows Agent B’s type
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy
   - $\text{mem}$: memory size
   - $\epsilon$: degree of randomness

2. Agent A knows Agent B’s type

   - Example: $\text{mem}=4$, $\epsilon = 0.1$
   - Agent A previous actions: $a_1, a_0, a_1, a_1$

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

© 2009 Peter Stone
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy.
   - $mem$: memory size
   - $\epsilon$: degree of randomness

2. Agent A knows Agent B’s type

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

   - Example: $mem= 4, \epsilon = 0.1$
     - Agent A previous actions: $a_1, a_0, a_1, a_1$
     - Agent B: A will select $a_0$ (prob. 0.25) or $a_1$ (0.75)
Assumptions

1. Agent B: bounded-memory BR, $\epsilon$-greedy action strategy.
   - \textit{mem}: memory size
   - $\epsilon$: degree of randomness

2. Agent A knows Agent B’s type

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Example: $\textit{mem} = 4$, $\epsilon = 0.1$
  - Agent A previous actions: $a_1, a_0, a_1, a_1$
  - Agent B: A will select $a_0$ (prob. 0.25) or $a_1$ (0.75)
  - $\text{BR}(a_1, a_0, a_1, a_1) = b_1$
Assumptions

1. Agent B: bounded-memory BR, \( \epsilon \)-greedy action strategy.
   - \textit{mem}: memory size
   - \textit{\( \epsilon \)}: degree of randomness

2. Agent A knows Agent B’s type

<table>
<thead>
<tr>
<th>( M )</th>
<th>( b_0 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

- Example: \( \text{mem} = 4, \epsilon = 0.1 \)
  - Agent A previous actions: \( a_1, a_0, a_1, a_1 \)
  - Agent B: A will select \( a_0 \) (prob. 0.25) or \( a_1 \) (0.75)
  - \( BR(a_1, a_0, a_1, a_1) = b_1 \)
  - Agent B: selects \( b_1 \) (1-\( \epsilon \)) or uniformly random (\( \epsilon \))
Assumptions

1. **Agent B**: bounded-memory BR, $\epsilon$-greedy action strategy.
   - $\text{mem}$: memory size
   - $\epsilon$: degree of randomness

2. **Agent A** knows **Agent B**'s type

- **Example**: $\text{mem} = 4$, $\epsilon = 0.1$
  - **Agent A** previous actions: $a_1, a_0, a_1, a_1$
  - **Agent B**: A will select $a_0$ (prob. 0.25) or $a_1$ (0.75)
  - $\text{BR}(a_1, a_0, a_1, a_1) = b_1$
  - **Agent B**: selects $b_1 (1-\epsilon)$ or uniformly random ($\epsilon$)
  - **Agent A**: action determines payoff and next history

<table>
<thead>
<tr>
<th>$M_1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Extensive Form Version

\[
\begin{array}{c}
\text{a1;0;1;1} \\
\text{(a0,b0)} \\
\text{a0;1;0;1} \\
\text{(a0,b0)} \\
\text{a0;0;1;0} \\
\text{a1;0;1;0} \\
\text{a2;0;1;0}
\end{array}
\]
Extensive Form Version

Stick with iterated normal form for presentation, algorithms
Questions

• Can we find the optimal action sequence efficiently?
• How long can the optimal action sequences be?

Cases

• Deterministic teammate, 1-step memory \((\text{mem}=1, \epsilon = 0)\)
• Longer teammate memory \((\text{mem}>1, \epsilon = 0)\)
• Teammate non-determinism \((\text{mem}>1, \epsilon > 0)\)
Dynamic Programming Algorithm

- Define $S^*_n(a_i, b_j) = \text{optimal sequence of length } n$
- Define $S^*_0(a_i, b_j)$ to be cost 0 if $m_{i,j} = m_*$, else $\infty$

\[ S^*_0(a_2, b_2) \text{ Cost 0} \]

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Dynamic Programming Algorithm

- Define $S^*_n(a_i, b_j) = \text{optimal sequence of length } n$
- Define $S^*_0(a_i, b_j)$ to be cost 0 if $m_{i,j} = m_*$, else $\infty$

<table>
<thead>
<tr>
<th></th>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^*_0(a_2, b_2)$</td>
<td>Cost 0</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$S^*_2(a_0, b_0)$</td>
<td>Cost 15+40</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$S^*_2(a_1, b_0)$</td>
<td>Cost 30+7</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
<tr>
<td>$S^*_2(a_2, b_0)$</td>
<td>Cost 40</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Dynamic Programming Algorithm

- Define $S^*_n(a_i, b_j) = \text{optimal sequence of length } n$
- Define $S^*_0(a_i, b_j)$ to be cost 0 if $m_{i,j} = m_*$, else $\infty$
- Find $S^*_n(a_i, b_j)$ using $S^*_{n-1}$'s $(O(d), d = \text{dim}(M))$
  - Either $S^*_{n-1}(a_i, b_j)$ or
  - Best sequence that prepends $(a_i, b_j)$ to $S^*_{n-1}(a_{act}, b_{BR}(a_i))$

<table>
<thead>
<tr>
<th>$S^*_3(a_0, b_0)$</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^*_2(a_0, b_0)$</td>
<td>Cost 55</td>
</tr>
<tr>
<td>$S^*_2(a_1, b_0)$</td>
<td>Cost 37</td>
</tr>
<tr>
<td>$S^*_2(a_2, b_0)$</td>
<td>Cost 40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Dynamic Programming Algorithm

- Define $S^*_n(a_i, b_j)$ = optimal sequence of length $n$
- Define $S^*_0(a_i, b_j)$ to be cost 0 if $m_{i,j} = m_*$, else $\infty$
- Find $S^*_n(a_i, b_j)$ using $S^*_{n-1}$'s $(O(d), d = \dim(M))$
  - Either $S^*_{n-1}(a_i, b_j)$ or
  - Best sequence that prepends $(a_i, b_j)$ to $S^*_{n-1}(a_{act}, b_{BR}(a_i))$

\[
\begin{align*}
S^*_3(a_0, b_0) & \quad \text{Cost 52} \\
S^*_2(a_0, b_0) & \quad \text{Cost 55} \\
S^*_2(a_1, b_0) & \quad \text{Cost 37+15} \\
S^*_2(a_2, b_0) & \quad \text{Cost 40+15}
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Dynamic Programming Algorithm

- Define $S_n^*(a_i, b_j) = \text{optimal sequence of length } n$
- Define $S_0^*(a_i, b_j)$ to be cost 0 if $m_{i,j} = m_*$, else $\infty$
- Find $S_n^*(a_i, b_j)$ using $S_{n-1}^*$'s
  - Either $S_{n-1}^*(a_i, b_j)$ or
  - Best sequence that prepends $(a_i, b_j)$ to $S_{n-1}^*(a_{act}, b_{BR(a_i)})$
- Sufficient to calculate $S_n^*(a_i, b_0), \forall i < x$
  - How high do we need to let $n$ get?

<table>
<thead>
<tr>
<th>$S_3^*(a_0, b_0)$</th>
<th>Cost 52</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2^*(a_0, b_0)$</td>
<td>Cost 55</td>
</tr>
<tr>
<td>$S_2^*(a_1, b_0)$</td>
<td>Cost 37+15</td>
</tr>
<tr>
<td>$S_2^*(a_2, b_0)$</td>
<td>Cost 40+15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M1$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>10</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>
Maximal Sequence Length

- Recall: Agent A has $x$ actions, Agent B has $y$

- **Theorem:** No sequence is longer than $\min(x, y)$
Maximal Sequence Length

- Recall: \textit{Agent A has} $x$ \textit{actions, Agent B has} $y$

- \textbf{Theorem}: No sequence is longer than $\min(x, y)$
  - Neither agent takes the same action twice
  - Otherwise, part of the sequence could be excised
Maximal Sequence Length

- Recall: Agent A has $x$ actions, Agent B has $y$

- **Theorem:** No sequence is longer than $\min(x, y)$
  - Neither agent takes the same action twice
  - Otherwise, part of the sequence could be excised

- **Theorem:** $\exists M$ with optimal sequence $\min(x, y)$
Maximal Sequence Length

- Recall: Agent A has \( x \) actions, Agent B has \( y \)

- **Theorem:** No sequence is longer than \( \min(x, y) \)
  - Neither agent takes the same action twice
  - Otherwise, part of the sequence could be excised

- **Theorem:** \( \exists M \) with optimal sequence \( \min(x, y) \)

<table>
<thead>
<tr>
<th>( M2 )</th>
<th>( b_0 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( \cdots )</th>
<th>( b_{y-3} )</th>
<th>( b_{y-2} )</th>
<th>( b_{y-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>( 100 - \delta )</td>
<td>0</td>
<td>0</td>
<td>( \cdots )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( 100 - 2\delta )</td>
<td>( 100 - \delta )</td>
<td>0</td>
<td>( \vdots )</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>0</td>
<td>( 100 - 2\delta )</td>
<td>( 100 - \delta )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td></td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a_{x-3} )</td>
<td>0</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( 100 - \delta )</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( a_{x-2} )</td>
<td>0</td>
<td>0</td>
<td>( \vdots )</td>
<td>( 100 - 2\delta )</td>
<td>( 100 - \delta )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( a_{x-1} )</td>
<td>D</td>
<td>0</td>
<td>0</td>
<td>( \cdots )</td>
<td>0</td>
<td>100</td>
<td>( 100 - 2\delta )</td>
</tr>
</tbody>
</table>
Maximal Sequence Length

• Recall: Agent A has $x$ actions, Agent B has $y$

• **Theorem:** No sequence is longer than $\min(x, y)$
  - Neither agent takes the same action twice
  - Otherwise, part of the sequence could be excised

• **Theorem:** $\exists M$ with optimal sequence $\min(x, y)$

<table>
<thead>
<tr>
<th>$M2$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$\cdots$</th>
<th>$b_{y-3}$</th>
<th>$b_{y-2}$</th>
<th>$b_{y-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>$100 - \delta$</td>
<td>0</td>
<td>0</td>
<td>$\cdots$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>$100 - 2\delta$</td>
<td>$100 - \delta$</td>
<td>0</td>
<td>$\vdots$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0</td>
<td>$100 - 2\delta$</td>
<td>$100 - \delta$</td>
<td>$\vdots$</td>
<td>0</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ddots$</td>
<td>$\ddots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$a_{x-3}$</td>
<td>0</td>
<td>$\vdots$</td>
<td>$\ddots$</td>
<td>$100 - \delta$</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$a_{x-2}$</td>
<td>0</td>
<td>0</td>
<td>$\vdots$</td>
<td>$100 - 2\delta$</td>
<td>$100 - \delta$</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$a_{x-1}$</td>
<td>D</td>
<td>0</td>
<td>0</td>
<td>$\cdots$</td>
<td>0</td>
<td>$100 - 2\delta$</td>
<td>100</td>
</tr>
</tbody>
</table>
Questions

- Find the optimal action sequence efficiently? \( O(d^3) \)
- Maximum length of optimal sequences? \( \min(x, y) \)

Cases

- Deterministic teammate, 1-step memory \((\text{mem} = 1, \epsilon = 0)\)
- Longer teammate memory \((\text{mem} > 1, \epsilon = 0)\)
- Teammate non-determinism \((\text{mem} > 1, \epsilon > 0)\)
Questions

- Find the optimal action sequence efficiently?
- Maximum length of optimal sequences?

Cases

- Deterministic teammate, 1-step memory \((\text{mem}=1, \epsilon = 0)\)
- Longer teammate memory \((\text{mem}>1, \epsilon = 0)\)
- Teammate non-determinism \((\text{mem}>1, \epsilon > 0)\)
Longer Teammate Memory

- Algorithm extends naturally, but exponential in $\text{mem}$
  - Need $S^*_{n-1}$ for every possible history of Agent A actions
  - Reaching $m^*$ once not sufficient ("stability")
Longer Teammate Memory

- Algorithm extends naturally, but exponential in $mem$
  - Need $S^*_{n-1}$ for every possible history of Agent A actions
  - Reaching $m^*$ once not sufficient ("stability")

<table>
<thead>
<tr>
<th>History</th>
<th>$[a_2; a_1; a_0]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M_3$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>0</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>$a_1$</td>
<td>41</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>$a_2$</td>
<td>99</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>
Longer Teammate Memory

- Algorithm extends naturally, but exponential in \( \text{mem} \)
  
  - Need \( S^*_{n-1} \) for every possible history of Agent A actions
  
  - Reaching \( m^* \) once not sufficient ("stability")

<table>
<thead>
<tr>
<th>History</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>([a_2; a_2; a_1])</td>
<td>(b_0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( M3 )</th>
<th>( b_0 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>0</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>41</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>99</td>
<td>20</td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
Longer Teammate Memory

- Algorithm extends naturally, but exponential in $\text{mem}$
  - Need $S^*_n$ for every possible history of Agent A actions
  - Reaching $m^*$ once not sufficient ("stability")

<table>
<thead>
<tr>
<th>History</th>
<th>$[a_2; a_2; a_2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M3$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>0</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>$a_1$</td>
<td>41</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>$a_2$</td>
<td>99</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>
Longer Teammate Memory

- Algorithm extends naturally, but exponential in $\text{mem}$
  - Need $S^*_{n-1}$ for every possible history of Agent A actions
  - Reaching $m^*$ once not sufficient (“stability”)

<table>
<thead>
<tr>
<th>History</th>
<th>$a_2; a_2; a_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

$$
\begin{array}{|c|c|c|c|}
\hline
M3 & b_0 & b_1 & b_2 \\
\hline
a_0 & 0 & 30 & 50 \\
a_1 & 41 & 20 & 0 \\
a_2 & 99 & 20 & 100 \\
\hline
\end{array}
$$

- NP-hard: reduction from Hamiltonian Path (Littman)
Longer Teammate Memory

\begin{itemize}
  \item \textbf{Theorem:} \( \exists M \) with optimal seq. \( (\min(x, y) - 1) \times \text{mem} + 1 \)
  \item \textbf{Conjecture:} No seq. longer than \( (\min(x, y) - 1) \times \text{mem} + 1 \)
\end{itemize}
Longer Teammate Memory

• **Theorem:** \( \exists M \text{ with optimal seq. } (\min(x, y) - 1) \ast \text{mem} + 1 \)

• **Conjecture:** No seq. longer than \( (\min(x, y) - 1) \ast \text{mem} + 1 \)
  – Can only prove no seq. longer than \( \min(x, y) \ast x^{\text{mem} - 1} \)
Longer Teammate Memory

- **Theorem:** \( \exists M \) with optimal seq. \((\min(x, y) - 1) \times \text{mem} + 1\)

- **Conjecture:** No seq. longer than \((\min(x, y) - 1) \times \text{mem} + 1\)
  - Can only prove no seq. longer than \(\min(x, y) \times x^{\text{mem}-1}\)

\[
\begin{array}{cccccccc}
M2 & b_0 & b_1 & b_2 & \cdots & b_{y-3} & b_{y-2} & b_{y-1} \\
\hline
a_0 & 100 - \delta & 0 & 0 & \cdots & 0 & 0 & 0 \\
a_1 & 100 - 2\delta & 100 - \delta & 0 & \vdots & 0 & 0 & 0 \\
a_2 & 0 & 100 - 2\delta & 100 - \delta & \vdots & 0 & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\
a_{x-3} & 0 & \vdots & \ddots & 100 - \delta & 0 & 0 \\
a_{x-2} & 0 & 0 & \vdots & 100 - 2\delta & 100 - \delta & 0 \\
a_{x-1} & 0 & 0 & 0 & \cdots & 0 & 100 - 2\delta & 100 \\
\end{array}
\]
Questions

• Find the optimal action sequence efficiently? no
• Maximum length of optimal sequences? ?

Cases

• Deterministic teammate, 1-step memory \((mem = 1, \epsilon = 0)\)
• Longer teammate memory \((mem > 1, \epsilon = 0)\)
• Teammate non-determinism \((mem > 1, \epsilon > 0)\)
Questions

- Find the optimal action sequence efficiently?
- Maximum length of optimal sequences?

Cases

- Deterministic teammate, 1-step memory \((\text{mem} = 1, \epsilon = 0)\)
- Longer teammate memory \((\text{mem} > 1, \epsilon = 0)\)
- Teammate non-determinism \((\text{mem} > 1, \epsilon > 0)\)
Teammate Non-Determinism

• \( EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k}) \)
  – Cost now sum of \( m^* - EV(a_i, b_j) \) over sequence
Teammate Non-Determinism

- \( EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k}) \)
  - Cost now sum of \( m^* - EV(a_i, b_j) \) over sequence
  - \( m^* \) now maximum \( EV(a_i, b_j) \) in \( M \)
Teammate Non-Determinism

- \( EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k}) \)
  - Cost now sum of \( m^* - EV(a_i, b_j) \) over sequence
  - \( m^* \) now maximum \( EV(a_i, b_j) \) in \( M \)

- “Target” (\( m^* \)) can change:
Teammate Non-Determinism

- $EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k})$
  - Cost now sum of $m^* - EV(a_i, b_j)$ over sequence
  - $m^*$ now maximum $EV(a_i, b_j)$ in $M$

- “Target” ($m^*$) can change: $S^*(a_0, b_0)$ with mem=3
Teammate Non-Determinism

• $EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k})$
  - Cost now sum of $m^* - EV(a_i, b_j)$ over sequence
  - $m^*$ now maximum $EV(a_i, b_j)$ in $M$

• “Target” ($m^*$) can change: $S^*(a_0, b_0)$ with mem=3

<table>
<thead>
<tr>
<th>$\epsilon$ = 0:</th>
<th>$m^*$ at ($a_3, b_3$)</th>
<th>$L(S^*)$ = 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$ = 0.1:</td>
<td>$m^*$ at ($a_3, b_3$)</td>
<td>$L(S^*)$ = 8</td>
</tr>
<tr>
<td>$\epsilon$ = 0.3:</td>
<td>$m^*$ at ($a_3, b_3$)</td>
<td>$L(S^*)$ = 3</td>
</tr>
<tr>
<td>$\epsilon$ = 0.4:</td>
<td>$m^*$ at ($a_2, b_2$)</td>
<td>$L(S^*)$ = 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M4$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>88</td>
<td>90</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_2$</td>
<td>70</td>
<td>98</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_3$</td>
<td>70</td>
<td>70</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>
Teammate Non-Determinism

- $EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k})$
  - Cost now sum of $m^* - EV(a_i, b_j)$ over sequence
  - $m^*$ now maximum $EV(a_i, b_j)$ in $M$

- “Target” ($m^*$) can change: $S^*(a_0, b_0)$ with $mem=3$

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$m^*$ at</th>
<th>$L(S^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$(a_3, b_3)$</td>
<td>10</td>
</tr>
<tr>
<td>0.1</td>
<td>$(a_3, b_3)$</td>
<td>8</td>
</tr>
<tr>
<td>0.3</td>
<td>$(a_3, b_3)$</td>
<td>3</td>
</tr>
<tr>
<td>0.4</td>
<td>$(a_2, b_2)$</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$M4$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>88</td>
<td>90</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_2$</td>
<td>70</td>
<td>98</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_3$</td>
<td>70</td>
<td>70</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>

- Algorithm and theorems hold unchanged
**Teammate Non-Determinism**

- $EV(a_i, b_j) = (1 - \epsilon)m_{i,j} + \frac{\epsilon}{y}(\sum_{k=0}^{y-1} m_{i,k})$
  - Cost now sum of $m^* - EV(a_i, b_j)$ over sequence
  - $m^*$ now maximum $EV(a_i, b_j)$ in $M$

- “Target” ($m^*$) can change: $S^*(a_0, b_0)$ with $\text{mem}=3$
  
  \[
  \begin{array}{|c|c|c|c|c|}
  \hline
  \epsilon = 0: & m^* \text{ at } (a_3, b_3) & L(S^*)=10 \\
  \hline
  \epsilon = 0.1: & m^* \text{ at } (a_3, b_3) & L(S^*)=8 \\
  \hline
  \epsilon = 0.3: & m^* \text{ at } (a_3, b_3) & L(S^*)=3 \\
  \hline
  \epsilon = 0.4: & m^* \text{ at } (a_2, b_2) & L(S^*)=3 \\
  \hline
  \end{array}
  \]

- Algorithm and theorems hold unchanged
  - Except when $\epsilon = 1$

<table>
<thead>
<tr>
<th>$M4$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$a_1$</td>
<td>88</td>
<td>90</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_2$</td>
<td>70</td>
<td>98</td>
<td>99</td>
<td>80</td>
</tr>
<tr>
<td>$a_3$</td>
<td>70</td>
<td>70</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>
Questions

- Find the optimal action sequence efficiently? no
- Maximum length of optimal sequences? ?

Cases

- Deterministic teammate, 1-step memory \((mem = 1, \epsilon = 0)\)
- Longer teammate memory \((mem > 1, \epsilon = 0)\)
- Teammate non-determinism \((mem > 1, \epsilon > 0)\)
Experiments

- All variations of the algorithm fully implemented
Experiments

- All variations of the algorithm fully implemented
- Test frequency of longest $S^*$ of varying lengths
  - 3x3 matrix: how often $L(S^*(a_i, b_j)) = 3$?
Experiments

- All variations of the algorithm fully implemented
- Test frequency of longest $S^*$ of varying lengths
  - 3x3 matrix: how often $L(S^*(a_i, b_j)) = 3$?
- $m_{i,j}$ uniformly random in $[0, 100]$; $m_{x-1,y-1} = 100$
Experiments

- All variations of the algorithm fully implemented
- Test frequency of longest $S^*$ of varying lengths
  - 3x3 matrix: how often $L(S^*(a_i, b_j)) = 3$?
- $m_{i,j}$ uniformly random in $[0, 100]$; $m_{x-1,y-1} = 100$

<table>
<thead>
<tr>
<th>mem=1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 x 3</td>
<td>104</td>
<td>852</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- All variations of the algorithm fully implemented
- Test frequency of longest $S^*$ of varying lengths
  - 3x3 matrix: how often $L(S^*(a_i, b_j)) = 3$?
- $m_{i,j}$ uniformly random in $[0, 100]$; $m_{x-1,y-1} = 100$

<table>
<thead>
<tr>
<th>mem=1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 × 3</td>
<td>104</td>
<td>852</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 × 4</td>
<td>12</td>
<td>825</td>
<td>158</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 × 5</td>
<td>3</td>
<td>662</td>
<td>316</td>
<td>19</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 × 6</td>
<td>0</td>
<td>465</td>
<td>489</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 × 7</td>
<td>0</td>
<td>349</td>
<td>565</td>
<td>81</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 × 8</td>
<td>0</td>
<td>236</td>
<td>596</td>
<td>159</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 × 9</td>
<td>0</td>
<td>145</td>
<td>640</td>
<td>193</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10 × 10</td>
<td>0</td>
<td>72</td>
<td>636</td>
<td>263</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Experiments

\[
\begin{array}{|c|cccccccccc|}
\hline
\text{mem}=1 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hline
3 \times 3 & 104 & 852 & 44 & & & & & & & \\
4 \times 4 & 12 & 825 & 158 & 5 & & & & & & \\
5 \times 5 & 3 & 662 & 316 & 19 & 0 & & & & & \\
6 \times 6 & 0 & 465 & 489 & 45 & 1 & 0 & & & & \\
7 \times 7 & 0 & 349 & 565 & 81 & 5 & 0 & 0 & & & \\
8 \times 8 & 0 & 236 & 596 & 159 & 8 & 1 & 0 & 0 & & \\
9 \times 9 & 0 & 145 & 640 & 193 & 20 & 2 & 0 & 0 & 0 & \\
10 \times 10 & 0 & 72 & 636 & 263 & 29 & 0 & 0 & 0 & 0 & 0 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|ccccccccccc|}
\hline
\text{mem}=3 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
\hline
3 \times 3 & 98 & 178 & 344 & 340 & 28 & 8 & 4 & 0 & 0 & 0 & 0 \\
4 \times 4 & 15 & 76 & 266 & 428 & 134 & 60 & 21 & 0 & 0 & 0 & 0 \\
5 \times 5 & 1 & 19 & 115 & 408 & 234 & 145 & 71 & 7 & 0 & 0 & 0 \\
6 \times 6 & 0 & 0 & 22 & 282 & 272 & 222 & 164 & 27 & 11 & 0 & 0 \\
7 \times 7 & 0 & 0 & 5 & 116 & 293 & 282 & 220 & 55 & 17 & 10 & 1 \\
\hline
\end{array}
\]
Robot Experiments

- In progress...
Related Work

Game Theory

- **Multiagent learning** (Claus & Boutilier, ’98), (Littman, ’01), (Conitzer & Sandholm, ’03), (Powers & Shoham, ’05), (Chakraborty & Stone, ’08)
- **Economic repeated games** (Hart & Mas-Colell, ’00), (Neyman & Okada, ’00)
- **Fictitious play** (Brown, ’51)
- **Adaptive play** (Young, ’93)

Opponent Modeling

- **Intended plan recognition** (Sidner, ’85), (Lochbaum, ’91), (Carberry, ’01)
- **SharedPlans** (Grosz & Kraus, ’96)
- **Recursive Modeling** (Vidal & Durfee, ’95)
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

Goal: Create a good team player

- Minimal representative scenarios
  - One teammate, no communication
  - Fixed and known behavior: best response
Scenarios

- Cooperative normal form game (w/ Kaminka & Rosenschein)

\[
\begin{array}{|c|c|c|c|}
\hline
M1 & b_0 & b_1 & b_2 \\
\hline
a_0 & 25 & 1 & 0 \\
a_1 & 10 & 30 & 10 \\
a_2 & 0 & 33 & 40 \\
\hline
\end{array}
\]

- Cooperative $k$-armed bandit (w/ Kraus)
3-armed bandit

- Random value from a distribution
- Expected value $\mu$
3-armed bandit

Arm∗  Arm₁  Arm₂
3-armed bandit

- Agent A: teacher
  - Knows payoff distributions
  - Objective: maximize expected sum of payoffs
3-armed bandit

- **Agent A: teacher**
  - Knows payoff distributions
  - Objective: maximize expected sum of payoffs
  - If alone, always Arm$_*$

\[ \mu_\ast > \mu_1 > \mu_2 \]
3-armed bandit

- Agent A: teacher
  - Knows payoff distributions
  - Objective: maximize expected sum of payoffs
  - If alone, always Arm
- Agent B: learner
  - Can only pull Arm or Arm

\[ \mu_\star > \mu_1 > \mu_2 \]
3-armed bandit

- **Agent A**: teacher
  - Knows payoff distributions
  - Objective: maximize expected sum of payoffs
  - If alone, always Arm$_*$

- **Agent B**: learner
  - Can only pull Arm$_1$ or Arm$_2$
  - Selects arm with highest observed sample average

$\mu_*$ > $\mu_1$ > $\mu_2$
Assumptions

$\text{Arm}_*$  \hspace{1cm}  $\text{Arm}_1$  \hspace{1cm}  $\text{Arm}_2$
Assumptions

- Alternate actions (teacher first)
- Results of all actions fully observable (to both)
Assumptions

- Alternate actions (teacher first)
- Results of all actions fully observable (to both)
- Number of rounds remaining finite, known to teacher
Assumptions

- Alternate actions (teacher first)
- Results of all actions fully observable (to both)
- Number of rounds remaining finite, known to teacher

Objective: maximize expected sum of payoffs
Formalism

• $\mu_i$: expected payoff of Arm$_i$ ($i \in \{1, 2, *\}$)

  – Assume $\mu_* > \mu_1 > \mu_2$: only interesting case
Formalism

- $\mu_i$: expected payoff of Arm$_i$ ($i \in \{1, 2, *\}$)
  
  - Assume $\mu_* > \mu_1 > \mu_2$: only interesting case

- $n_i$: number of times Arm$_i$ has been pulled

- $m_i$: cumulative payoff from past pulls of Arm$_i$
Formalism

- $\mu_i$: expected payoff of Arm$_i$ ($i \in \{1, 2, \ast\}$)
  - Assume $\mu_\ast > \mu_1 > \mu_2$: only interesting case

- $n_i$: number of times Arm$_i$ has been pulled

- $m_i$: cumulative payoff from past pulls of Arm$_i$

- $\bar{x}_i = \frac{m_i}{n_i}$: observed sample average so far
Formalism

• \( \mu_i \): expected payoff of Arm\(_i\) \( (i \in \{1, 2, *\}) \)
  - Assume \( \mu_* > \mu_1 > \mu_2 \): only interesting case

• \( n_i \): number of times Arm\(_i\) has been pulled

• \( m_i \): cumulative payoff from past pulls of Arm\(_i\)

• \( \bar{x}_i = \frac{m_i}{n_i} \): observed sample average so far

• \( r \): number of rounds left
Formalism

- $\mu_i$: expected payoff of Arm$_i$ ($i \in \{1, 2, *\}$)
  - Assume $\mu_* > \mu_1 > \mu_2$: only interesting case
- $n_i$: number of times Arm$_i$ has been pulled
- $m_i$: cumulative payoff from past pulls of Arm$_i$
- $\bar{x}_i = \frac{m_i}{n_i}$: observed sample average so far
- $r$: number of rounds left

Which arm should the teacher pull, as a function of $r$ and all the $\mu_i$, $n_i$, and $\bar{x}_i$?

© 2009 Peter Stone
Teacher should consider Arm$_1$

$\mu_\ast = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \quad r = 3$
Teacher should consider Arm_1

\[ \mu_* = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \]

\[ r = 3 \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>( m_i )</th>
<th>( n_i )</th>
<th>( \bar{x}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

9.8
Teacher should consider Arm$_1$

$\mu_* = 10.0$  \hspace{1cm} $\mu_1 = 9.0$  \hspace{1cm} $\mu_2 = 5.0$

$r = 3$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

$9.8$  \hspace{1cm} $7.0$
Teacher should consider Arm\textsubscript{1}

\[
\begin{align*}
\mu_\ast &= 10.0 \\
\mu_1 &= 9.0 \\
\mu_2 &= 5.0 \\
r &= 2
\end{align*}
\]

\[
\begin{array}{ccc}
 i & m_i & n_i \\
 1 & 0.0 & 0 \\
 2 & 7.0 & 1 \\
\end{array}
\]
Teacher should consider Arm$_1$

$\mu_\star = 10.0$  $\mu_1 = 9.0$  $\mu_2 = 5.0$  

$r = 2$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>
Teacher should consider Arm_1

\[ \mu_\star = 10.0 \]
\[ \mu_1 = 9.0 \]
\[ \mu_2 = 5.0 \]

\[ r = 2 \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>( m_i )</th>
<th>( n_i )</th>
<th>( \bar{x}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

10.3

6.0
Teacher should consider Arm_1

\[ \mu_* = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
<td></td>
</tr>
</tbody>
</table>
Teacher should consider $\text{Arm}_1$

$\mu_* = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \quad r = 1$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher $\text{Arm}_1$ expected value:
Teacher should consider $\text{Arm}_1$

$\mu_* = 10.0$, $\mu_1 = 9.0$, $\mu_2 = 5.0$

$r = 1$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher $\text{Arm}_1$ expected value:
  - Define $\eta$: probability $\text{Arm}_1$ returns $> 8$
  - Assume: $\eta > \frac{1}{2}$
Teacher should consider Arm_1

\[ \mu_* = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>( m_i )</th>
<th>( n_i )</th>
<th>( \bar{x}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher Arm_1 expected value:
  - Define \( \eta \): probability Arm_1 returns > 8
  - Assume: \( \eta > \frac{1}{2} \)
  - EV: \( \mu_1 + \eta \mu_1 + (1 - \eta) \mu_2 \)
Teacher should consider Arm$_1$

$\mu_* = 10.0$ \hspace{1cm} $\mu_1 = 9.0$ \hspace{1cm} $\mu_2 = 5.0$

$r = 1$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher Arm$_1$ expected value:
  - Define $\eta$: probability Arm$_1$ returns $> 8$
  - Assume: $\eta > \frac{1}{2}$
  - EV: $\mu_1 + \eta \mu_1 + (1 - \eta) \mu_2 > 9 + \frac{9}{2} + \frac{5}{2} = 16$
Teacher should consider Arm$_1$

\[ \mu_* = 10.0 \quad \mu_1 = 9.0 \quad \mu_2 = 5.0 \]

\[ r = 1 \]

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher Arm$_1$ expected value:
  - Define $\eta$: probability Arm$_1$ returns $> 8$
  - Assume: $\eta > \frac{1}{2}$
  - EV: $\mu_1 + \eta \mu_1 + (1 - \eta) \mu_2 > 9 + \frac{9}{2} + \frac{5}{2} = 16$

- Teacher Arm$_*$ expected value:
  - EV: $\mu_* + \mu_2$
Teacher should consider $\text{Arm}_1$

$\mu_\ast = 10.0$  $\mu_1 = 9.0$  $\mu_2 = 5.0$

$r = 1$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$m_i$</th>
<th>$n_i$</th>
<th>$\bar{x}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>1</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Teacher $\text{Arm}_1$ expected value:
  - Define $\eta$: probability $\text{Arm}_1$ returns $> 8$
  - Assume: $\eta > \frac{1}{2}$
  - $\text{EV}: \mu_1 + \eta \mu_1 + (1 - \eta) \mu_2 > 9 + \frac{9}{2} + \frac{5}{2} = 16$

- Teacher $\text{Arm}_\ast$ expected value:
  - $\text{EV}: \mu_\ast + \mu_2 = 15$
Should teacher consider $\text{Arm}_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
Should teacher consider $\text{Arm}_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from $\text{Arm}_2$: $u_0, u_1, u_2, \ldots$
Should teacher consider Arm$_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from Arm$_2$: $u_0$

© 2009 Peter Stone
Should teacher consider \( \text{Arm}_2 \)?

- \( \bar{x}_1 < \bar{x}_2 \implies \text{no} \)
  - Sequence of values from \( \text{Arm}_2 \): \( u_0, u_1, u_2, \ldots \)
  - Optimal from \( \text{Arm}_2 \): \( u_0, a \)
Should teacher consider Arm$_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
Should teacher consider Arm$_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*$
Should teacher consider $\text{Arm}_2$?

- $x_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from $\text{Arm}_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from $\text{Arm}_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, u_0, \mu_*$
Should teacher consider Arm$_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, u_0, \mu_*, a$
Should teacher consider Arm$_2$?

• $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  
  – Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  – Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  – Also possible: $\mu_*, u_0, \mu_*, a, b, c, d, e, \ldots, w, x$
Should teacher consider $\text{Arm}_2$?

- $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  - Sequence of values from $\text{Arm}_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from $\text{Arm}_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, u_0, \mu_*, a, b, c, d, e, \ldots, w, x$

- $\bar{x}_1 > \bar{x}_2 \implies ?$
Should teacher consider Arm$_2$?

• $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  
  – Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  
  – Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  
  – Also possible: $\mu_*, u_0, \mu_*, a, b, c, d, e, \ldots, w, x$

• $\bar{x}_1 > \bar{x}_2 \implies ?$

  – Subtle, but still no
Should teacher consider Arm$_2$?

• $\bar{x}_1 < \bar{x}_2 \implies \text{no}$
  
  – Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  – Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  – Also possible: $\mu_*, u_0, \mu_*, a, b, c, d, e, \ldots, w, x$

• $\bar{x}_1 > \bar{x}_2 \implies ?$
  
  – Subtle, but still no
  – Challenge: prove it!
Never teach when $\bar{x}_1 > \bar{x}_2$

- Same proof
Never teach when $\bar{x}_1 > \bar{x}_2$

- Same proof
  - Sequence of values from $\text{Arm}_1$: $v_0, v_1, v_2, \ldots$
  - Optimal from $\text{Arm}_1$: $v_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu^*, v_0, \mu^*, a, b, c, d, e, \ldots, w, x$
Never teach when $\bar{x}_1 > \bar{x}_2$

- Same proof
  - Sequence of values from Arm$_1$: $v_0, v_1, v_2, \ldots$
  - Optimal from Arm$_1$: $v_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu^*, v_0, \mu^*, a, b, c, d, e, \ldots, w, x$

- Only need to consider Arm$_1$ when $\bar{x}_1 < \bar{x}_2$
  - Depends on distributions
Never teach when $\bar{x}_1 > \bar{x}_2$

- Same proof
  - Sequence of values from Arm$_1$: $v_0, v_1, v_2, \ldots$
  - Optimal from Arm$_1$: $v_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, v_0, \mu_*, a, b, c, d, e, \ldots, w, x$

- Only need to consider Arm$_1$ when $\bar{x}_1 < \bar{x}_2$
  - Depends on distributions
  - Consider binary and normal
Arms with Binary Distributions

\[ \rightarrow \left\{ \begin{array}{ll} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{array} \right. \]
Arms with Binary Distributions

\[ \rightarrow \left\{ \begin{array}{ll} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{array} \right. \]

\[ \mu_i = p_i \quad m_i = \text{number of 1's so far} \]
Arms with Binary Distributions, \( r = 1 \)

\[
p_* > p_1 > p_2
\]
Arms with Binary Distributions, \( r = 1 \)

- Consider teaching if

  1. \( \bar{x}_1 < \bar{x}_2 \)
Arms with Binary Distributions, $r = 1$

- Consider teaching if

1. $\bar{x}_1 < \bar{x}_2 \quad \equiv \quad \frac{m_1}{n_1} < \frac{m_2}{n_2}$
Arms with Binary Distributions, $r = 1$

$\bar{x}_1 < \bar{x}_2 \equiv \frac{m_1}{n_1} < \frac{m_2}{n_2}$

1. $\bar{x}_1 < \bar{x}_2$  \[
\begin{align*}
p_* & > p_1, \quad p_1 > p_2
\end{align*}
\]

Consider teaching if

2. It could help: $\frac{m_1+1}{n_1+1} > \frac{m_2}{n_2}$
Arms with Binary Distributions, $r = 1$

- Consider teaching if
  1. $\bar{x}_1 < \bar{x}_2 \equiv \frac{m_1}{n_1} < \frac{m_2}{n_2}$
  2. It could help: $\frac{m_1+1}{n_1+1} > \frac{m_2}{n_2}$

- Teacher Arm $\star$ expected value: $p_\star + p_2$
Arms with Binary Distributions, \( r = 1 \)

\[
\begin{align*}
p_* & > p_1 & > & p_2 \\
\end{align*}
\]

- Consider teaching if
  
  1. \( \bar{x}_1 < \bar{x}_2 \) \( \equiv \) \( \frac{m_1}{n_1} < \frac{m_2}{n_2} \)
  
  2. It could help: \( \frac{m_1+1}{n_1+1} > \frac{m_2}{n_2} \)

- Teacher Arm* expected value: \( p_* + p_2 \)

- Teacher Arm1 expected value: \( p_1 \)
Arms with Binary Distributions, $r = 1$

- Consider teaching if:
  1. $\bar{x}_1 < \bar{x}_2 \equiv \frac{m_1}{n_1} < \frac{m_2}{n_2}$
  2. It could help: $\frac{m_1+1}{n_1+1} > \frac{m_2}{n_2}$

- Teacher Arm$_*$ expected value: $p_* + p_2$

- Teacher Arm$_1$ expected value: $p_1 + p_1 * p_1$
Arms with Binary Distributions, $r = 1$

- Consider teaching if
  1. $\bar{x}_1 < \bar{x}_2 \equiv \frac{m_1}{n_1} < \frac{m_2}{n_2}$
  2. It could help: $\frac{m_1+1}{n_1+1} > \frac{m_2}{n_2}$

- Teacher Arm$_*$ expected value: $p_* + p_2$

- Teacher Arm$_1$ expected value: $p_1 + p_1 \times p_1 + (1 - p_1)p_2$
Arms with Binary Distributions, \( r = 1 \)

\[
p_* > p_1 > p_2
\]

- Consider teaching if
  1. \( \bar{x}_1 < \bar{x}_2 \equiv \frac{m_1}{n_1} < \frac{m_2}{n_2} \)
  2. It could help: \( \frac{m_1+1}{n_1+1} > \frac{m_2}{n_2} \)

- Teacher Arm\( _* \) expected value: \( p_* + p_2 \)

- Teacher Arm\( _1 \) expected value: \( p_1 + p_1 \times p_1 + (1 - p_1)p_2 \)

Teach iff conditions 1, 2, and \( p_* - p_1 < p_1(p_1 - p_2) \)
Algorithm for Optimal Teacher Action

- Polynomial algorithm finds optimal teacher action
  - Takes starting values $M_1, N_1, M_2, N_2$ and $R$
Algorithm for Optimal Teacher Action

- Polynomial algorithm finds optimal teacher action
  - Takes starting values $M_1, N_1, M_2, N_2$ and $R$

- Dynamic programming
  - Works backwards from $r = 1$
  - Considers all reachable values of $m_1, n_1, m_2, n_2$
Algorithm for Optimal Teacher Action

• Polynomial algorithm finds optimal teacher action
  – Takes starting values $M_1, N_1, M_2, N_2$ and $R$

• Dynamic programming
  – Works backwards from $r = 1$
  – Considers all reachable values of $m_1, n_1, m_2, n_2$

• $O(r^5)$ in both memory and runtime
Arms with Normal Distributions

\[ \rightarrow N(\mu, \sigma) \]
Arms with Normal Distributions, $r = 1$

$(\mu_*, \sigma_*)$  
$(\mu_1, \sigma_1)$  
$(\mu_2, \sigma_2)$
Arms with Normal Distributions, $r = 1$

\[
(\mu_*, \sigma_*) \quad (\mu_1, \sigma_1) \quad (\mu_2, \sigma_2)
\]

- Cost of teaching: $\mu_* - \mu_1$

\[c\]
Arms with Normal Distributions, $r = 1$

- Cost of teaching: $\mu_* - \mu_1$
- Benefit of teaching if successful: $\mu_1 - \mu_2$
Arms with Normal Distributions, $r = 1$

$$(\mu_*, \sigma_*)$$  $$((\mu_1, \sigma_1))$$  $$(\mu_2, \sigma_2)$$

- Cost of teaching: $\mu_* - \mu_1$
- Benefit of teaching if successful: $\mu_1 - \mu_2$  \hspace{1cm} ($\bar{x}_1 < \bar{x}_2$)
Arms with Normal Distributions, $r = 1$

- Cost of teaching: $\mu_* - \mu_1$
- Benefit of teaching if successful: $\mu_1 - \mu_2$ \hspace{1cm} ($\bar{x}_1 < \bar{x}_2$)
- Probability it’s successful: $1 - \Phi_{\mu_1, \sigma_1}(\bar{x}_2(n_1 + 1) - \bar{x}_1 n_1)$
  - Cumulative probability that pulling Arm_1 causes $\bar{x}_1 > \bar{x}_2$
Arms with Normal Distributions, \( r = 1 \)

\[
(\mu_*, \sigma_*) \quad (\mu_1, \sigma_1) \quad (\mu_2, \sigma_2)
\]

- Cost of teaching: \( \mu_* - \mu_1 \)
- Benefit of teaching if successful: \( \mu_1 - \mu_2 \) \( (\bar{x}_1 < \bar{x}_2) \)
- Probability it’s successful: \( 1 - \Phi_{\mu_1, \sigma_1}(\bar{x}_2(n_1 + 1) - \bar{x}_1n_1) \)
  - Cumulative probability that pulling Arm_1 causes \( \bar{x}_1 > \bar{x}_2 \)

**Teach iff** \[
1 - \Phi_{\mu_1, \sigma_1}(\bar{x}_2(n_1 + 1) - \bar{x}_1n_1) > \frac{\mu_* - \mu_1}{\mu_1 - \mu_2}
\]
Arms with Normal Distributions, \( r \geq 2 \)

- Can solve computationally — nested integral
- Not exactly, nor efficiently
Arms with Normal Distributions, \( r \geq 2 \)

- Can solve computationally — nested integral
- Not exactly, nor efficiently
- Can you find an efficient algorithm?
Experiments

- Evaluating teacher heuristics
Experiments

- Evaluating teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x}_1 < \bar{x}_2$
  3. Teach iff it would be optimal to teach if $r = 1$
Experiments

- Evaluating teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x}_1 < \bar{x}_2$
  3. Teach iff it would be optimal to teach if $r = 1$
     - None dominates
Experiments

- Evaluating teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x}_1 < \bar{x}_2$
  3. Teach iff it would be optimal to teach if $r = 1$
     - None dominates

- Looking for patterns in optimal action as a function of $r$
Experiments

- Evaluating teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x}_1 < \bar{x}_2$
  3. Teach iff it would be optimal to teach if $r = 1$
     - None dominates

- Looking for patterns in optimal action as a function of $r$
  - Conjecture: teach when $r = 1 \implies$ teach when $r = 2$
Experiments

• Evaluating teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x}_1 < \bar{x}_2$
  3. Teach iff it would be optimal to teach if $r = 1$
     - None dominates

• Looking for patterns in optimal action as a function of $r$
  - Conjecture: teach when $r = 1 \implies$ teach when $r = 2$
  - False!
Experiments

● Evaluating teacher heuristics

1. Never teach
2. Teach iff $\overline{x}_1 < \overline{x}_2$
3. Teach iff it would be optimal to teach if $r = 1$
   - None dominates

● Looking for patterns in optimal action as a function of $r$
   - Conjecture: teach when $r = 1 \implies$ teach when $r = 2$
   - **False!** (binary and normal)
More than 3 arms

Arm\_1 \quad Arm\_2 \quad Arm\_3 \quad Arm_1 \quad Arm_2 \quad Arm_3 \quad \cdots \quad Arm_z
More than 3 arms

- Additional arms for teacher make no difference
More than 3 arms

- Additional arms for teacher make no difference
  - Ignore all but the best
More than 3 arms

- Additional arms for teacher make no difference
  - Ignore all but the best

- Additional learner arms: most results generalize naturally
Additional arms for teacher make no difference
  - Ignore all but the best

Additional learner arms: most results generalize naturally
  - Never teach with Arm\(_z\)
More than 3 arms

- Additional arms for teacher make no difference
  - Ignore all but the best

- Additional learner arms: most results generalize naturally
  - Never teach with $\text{Arm}_z$ ($\text{Arm}_1$–$\text{Arm}_{z-1}$ possible)
More than 3 arms

- Additional arms for teacher make no difference
  - Ignore all but the best

- Additional learner arms: most results generalize naturally
  - Never teach with Arm\(_z\) (Arm\(_1\)–Arm\(_{z-1}\) possible)
  - Never teach with Arm\(_i\) when \(\bar{x}_i > \bar{x}_j, \forall j \neq i\)
More than 3 arms

- Additional arms for teacher make no difference
  - Ignore all but the best

- Additional learner arms: most results generalize naturally
  - Never teach with Arm\textsubscript{z} (Arm\textsubscript{1}–Arm\textsubscript{z−1} possible)
  - Never teach with Arm\textsubscript{i} when $\bar{x}_i > \bar{x}_j, \forall j \neq i$
  - **Surprising:** May be best to teach with Arm\textsubscript{j} for $j > i$
More than 3 arms

<table>
<thead>
<tr>
<th>Arm*</th>
<th>Arm$_1$</th>
<th>Arm$_2$</th>
<th>Arm$_3$</th>
<th>\ldots</th>
<th>Arm$_z$</th>
</tr>
</thead>
</table>

- Additional arms for teacher make no difference
  - Ignore all but the best

- Additional learner arms: most results generalize naturally
  - Never teach with Arm$_z$ (Arm$_1$–Arm$_{z-1}$ possible)
  - Never teach with Arm$_i$ when $\bar{x}_i > \bar{x}_j$, $\forall j \neq i$
  - **Surprising:** May be best to teach with Arm$_j$ for $j > i$
    (teach with Arm$_2$, even though $\bar{x}_1 > \bar{x}_2 > \bar{x}_3$)
Sample Open Questions

- What if the teacher doesn’t know the distributions?
Sample Open Questions

- What if the teacher doesn’t know the distributions?
  - Exploration vs. exploitation
Sample Open Questions

- What if the teacher doesn’t know the distributions?
  - Exploration vs. exploitation vs. teaching
Sample Open Questions

- What if the teacher doesn’t know the distributions?
  - Exploration vs. exploitation vs. teaching

- What if the learner isn’t greedy: explores on its own?
Sample Open Questions

- What if the teacher doesn’t know the distributions?
  - Exploration vs. exploitation vs. teaching
- What if the learner isn’t greedy: explores on its own?
- How does this extend to the infinite (discounted) case?
Sample Open Questions

- What if the teacher doesn’t know the distributions?
  - Exploration vs. exploitation vs. teaching

- What if the learner isn’t greedy: explores on its own?

- How does this extend to the infinite (discounted) case?

- What if there are multiple learners?
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

**Goal:** Create a good team player
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

**Goal:** Create a good team player

- **So far:** Minimal representative scenarios
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

Goal: Create a good team player

- So far: Minimal representative scenarios
- Future: Unknown teammate behavior, communication, incomplete teacher knowledge, . . .
Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control

**Goal:** Create a good team player

- **So far:** Minimal representative scenarios
- **Future:** Unknown teammate behavior, communication, incomplete teacher knowledge, . . . **much more!**
Acknowledgements

- Yonatan Aumann, Michael Littman, Reshef Meir, Jeremy Stober, Daniel Stronger
- Fulbright and Guggenheim Foundations
- Israel Science Foundation