Emergent Communication for Collaborative Reinforcement Learning

Yarin Gal and Rowan McAllister

MLG RCC

8 May 2014

Game Theory

Multi-Agent Reinforcement Learning

Learning Communication

Nash Equilibrium

UNIVERSITY OF CAMBRIDGE

"**Nash equilibria** *are game-states s.t. no player would fare better by unilateral*[1] *change of their own action.*"

[1] Performed by or affecting only one person involved in a situation, without the agreement of another.

Prisoner's Dilemma

		Sideshow Bob	
		Cooperate	Defect
Snake	Cooperate	1,1	3,0
	Defect	0,3	2,2

(prison sentence in years)

Pareto Efficiency

"**Pareto optima** *are game-states s.t. no alternative state exists whereby each player would fare equal or better.*"

$\mathcal{A}_t = \{\text{Cooperate } (C), \text{Defect } (D)\}$

$\mathcal{S}_t = \{CC, CD, DC, DD\}$ (previous game outcome)

$\pi : \times_{i=2}^{t} \mathcal{S}_i \rightarrow \mathcal{A}_t$

Possible strategies π for Snake:

- Tit-for-Tat:

$$\pi(s_t) = \begin{cases} C & \text{, if } t = 1; \\ a_{Bob,t-1} & \text{, if } t > 1 \end{cases}$$

- Reinforce actions conditioned on game outcomes:

$$\pi(s_t) = \arg\min_a \mathbb{E}_{\mathcal{T}}[\text{accumulated prison years} | s_t, a]$$

update transition model $\mathcal{T}$

Game Theory

Multi-Agent Reinforcement Learning

Learning Communication

How can we learn mutually-beneficial collaboration strategies?

- **Modelling**:
  multi-agent-MDPs, dec-MDPs
- **Issues solving joint tasks**:
  - decentralised knowledge with no centralised control,
  - credit assignment,
  - communication constraints
- **Issues affecting individual agents**:
  - state space explodes: $\mathcal{O}(|\mathcal{S}|^{\#\textit{agents}})$,
  - coadapatation $\rightarrow$ dynamic non-Markov environment

Markov Decision Process (MDP)

Stochastic environment characterised by tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma\}$, where:

- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow \mathbb{R} \in (-\infty, \infty)$
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow \mathbb{R} \in [0, 1]$
- $\gamma \in [0, 1]$

Multi-agent MDP (MMDP)

N-agent stochastic game characterised by tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma\}$, where:

- $\mathcal{S} = \times_{i=1}^{N} \mathcal{S}_i$
- $\mathcal{A} = \times_{i=1}^{N} \mathcal{A}_i$
- $\mathcal{R} = \times_{i=1}^{N} \mathcal{R}_i, \qquad \mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow \mathbb{R}$
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow \mathbb{R}$

Multi-agent Q-learning

- Oblivious agents [Sen et al., 1994]

$$
\begin{aligned}
Q_i(s, a_i) &\leftarrow (1-\alpha) Q_i(s, a_i) + \alpha[\mathcal{R}_i(s, a_i) + \gamma V_i(s')] \\
V_i^*(s) &= \max_{a_i \in \mathcal{A}_i} Q_i^*(s, a_i)
\end{aligned}
$$

- Common-payoff games [Claus and Boutilier, 1998]

$$
\begin{aligned}
Q_i(s, a) &\leftarrow (1-\alpha) Q_i(s, a) + \alpha[\mathcal{R}_i(s, a, s') + \gamma V_i(s')] \\
V_i(s) &\leftarrow \max_{a_i \in \mathcal{A}} \sum_{a_{-i} \in \mathcal{A}/\{\mathcal{A}_i\}} P_i(s, a_{-i}) Q_i(s, \{a_i, a_{-i}\})
\end{aligned}
$$

Multi-agent Q-learning

- Oblivious agents [Sen et al., 1994]

$$\begin{aligned} Q_i(s, a_i) &\leftarrow (1-\alpha) Q_i(s, a_i) + \alpha[\mathcal{R}_i(s, a_i) + \gamma V_i(s')] \\ V_i^*(s) &= \max_{a_i \in \mathcal{A}_i} Q_i^*(s, a_i) \end{aligned}$$

- Common-payoff games [Claus and Boutilier, 1998]

$$\begin{aligned} Q_i(s, a) &\leftarrow (1-\alpha) Q_i(s, a) + \alpha[\mathcal{R}_i(s, a, s') + \gamma V_i(s')] \\ V_i(s) &\leftarrow \max_{a_i \in \mathcal{A}} \sum_{a_{-i} \in \mathcal{A}/\{\mathcal{A}_i\}} P_i(s, a_{-i}) Q_i(s, \{a_i, a_{-i}\}) \end{aligned}$$

Independent vs Cooperative Learning

[Tan, 1993]: "*Can N communicating agents outperform N non-communicating agents?*"

Ways of communication:

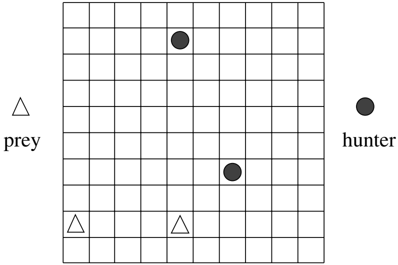
- Agents share Q-learning updates (thus syncing Q-values):
  - Pro: each agent learns N-fold faster (per timestep),
  - Note: same asymptotic performance as independent agents.
- Agents share sensory information:
  - Pro: more information $\rightarrow$ better policies,
  - Con: more information $\rightarrow$ larger state space $\rightarrow$ slower learning.

[Tan, 1993]: "*Can N communicating agents outperform N non-communicating agents?*"

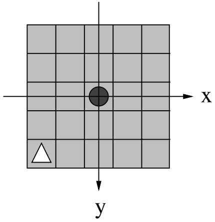
Ways of communication:

- Agents share Q-learning updates (thus syncing Q-values):
  - Pro: each agent learns N-fold faster (per timestep),
  - Note: same asymptotic performance as independent agents.
- Agents share sensory information:
  - Pro: more information $\rightarrow$ better policies,
  - Con: more information $\rightarrow$ larger state space $\rightarrow$ slower learning.

Hunter-Prey Problem



10×10 grid world.



perceptual state, visual depth 2 (prey's relative position).
$|\mathcal{S}| = 5^2 + 1 = 26$

$$\mathcal{R} = \begin{cases} 1.0 & : \text{ a hunter catches a prey, i.e. } \{x_i, y_i\} = \{0, 0\} \\ -0.1 & : \text{ otherwise} \end{cases}$$

Hunter-Prey Experiments

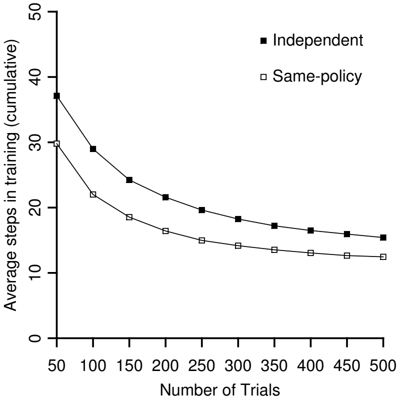
Experiment 1 – any hunter catches a prey:

- Baseline: 2 independent hunters, $|\mathcal{S}_i| = 5^2 + 1 = 26$
- 2 hunters, communicating Q-value updates. $|\mathcal{S}_i| = 26$

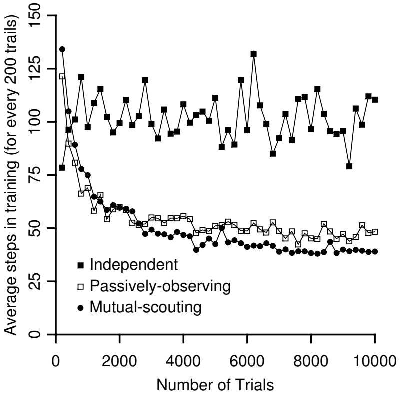
Experiment 2 – both hunters catch same prey simultaneously:

- Baseline: 2 independent hunters, $|\mathcal{S}_i| = 26$
- 2 hunters, communicating own locations, $|\mathcal{S}_i| \approx 26 \cdot 19^2 = 9386$
- 2 hunters, communicating own+prey locations. $|\mathcal{S}_i| \approx (19^2 + 1) \cdot 19^2 = 130682$

Hunter-Prey Results



Experiment 1: any hunter catches a prey



Experiment 2: Both hunters catching same prey simultaneously

UNIVERSITY OF CAMBRIDGE

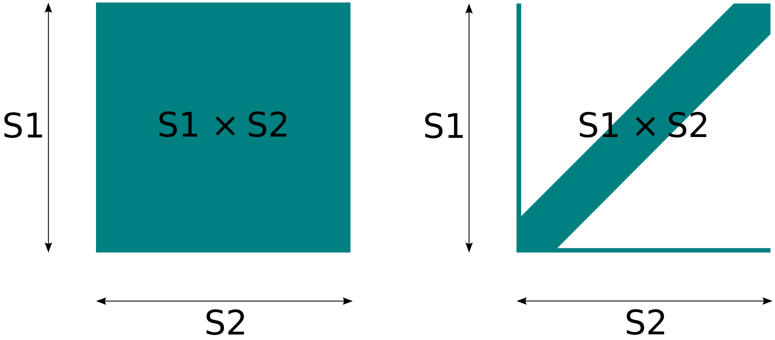
[Melo and Veloso, 2011] Philosophy:

- N-agent coordination is hard since the size of the state space grows exponentially in N.
- $\therefore$ Limit scope of coordination to where it's probably more useful; plans and learn w.r.t. 'local' agent-agent interactions only.

The Dec-SIMDP framework determines when and how agents i and j coordinate vs act independently.

'Decentralised' $=$ have full *joint* $\mathcal{S}$-observability, but not full *individual* $\mathcal{S}$-observability (agent i only observes $\mathcal{S}_i$ $+$ nearby agents).

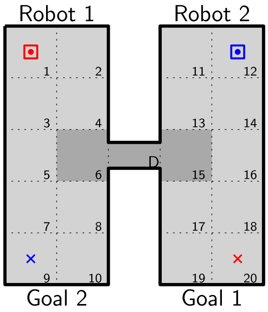
Dec-SIMDP: Reducing Joint State Space



Global coupling

Local coupling only

UNIVERSITY OF CAMBRIDGE

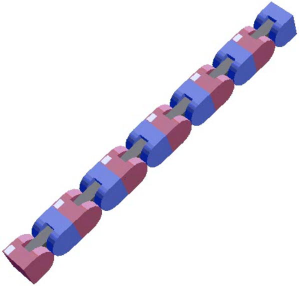


Navigation task: coordination necessarily only when crossing the narrow doorway.

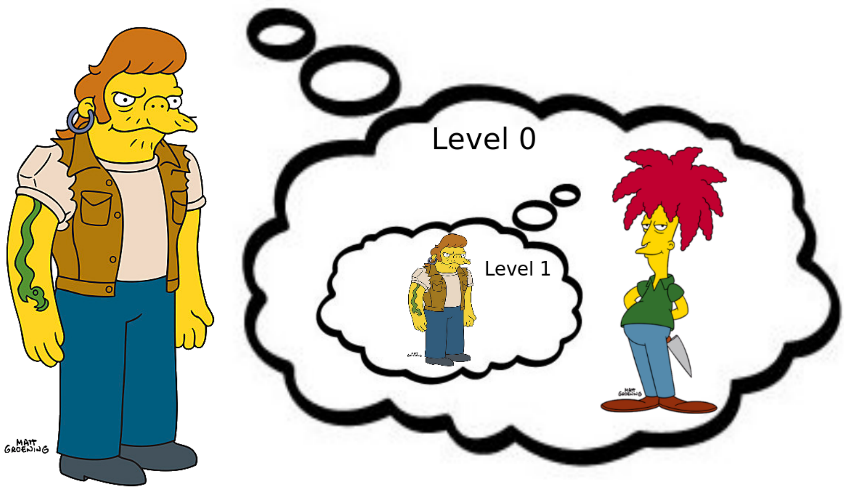
$$\mathcal{S}_i = \{1, ..., 20, D\},$$
$$\mathcal{A}_i = \{N, S, E, W\},$$
$$\mathcal{Z}_i = \mathcal{S}_i \cup \{\{6, 15, D\} \times \{6, 15, D\}\}$$

$$\mathcal{R}(s, a) = \begin{cases} 2 & \text{if } s = (20, 9) \\ 1 & \text{if } s_1 = 20, \text{ or } s_2 = 9 \\ -20 & \text{if } s = (D, D) \\ 0 & \text{otherwise} \end{cases}$$

Four interconnected modular robots cooperate to change configuration: line $\rightarrow$ ring



[Mundhe and Sen, 2000]



How should individuals be individually credited w.r.t. total team performance (or utility)?

$\leftrightarrow$

Communication

"Shall we both choose to cooperate next round?"

Snake		**Sideshow Bob**: Cooperate	**Sideshow Bob**: Defect
	Cooperate	1,1	3,0
	Defect	0,3	2,2

(prison sentence in years)

"OK."

Alien

Snake	Cooperate	Defect
Cooperate	1,1	3,0
Defect	0,3	2,2

(prison sentence in years)

“*What?*”

Game Theory

Multi-Agent Reinforcement Learning

Learning Communication

UNIVERSITY OF CAMBRIDGE

- How *learning* communication can help in RL collaboration
- Approaches to learning communication (ranging from linguistically motivated to a pragmatic view)
- What problems exist with learning communication?

How can learning communication help in RL collaboration?

- Forgoes expensive expert time for protocol planning
- Allows for a decentralised system without an external authority to decide on a communication protocol
- Life-long learning (adaptive tasks, e.g. future proofed robots)

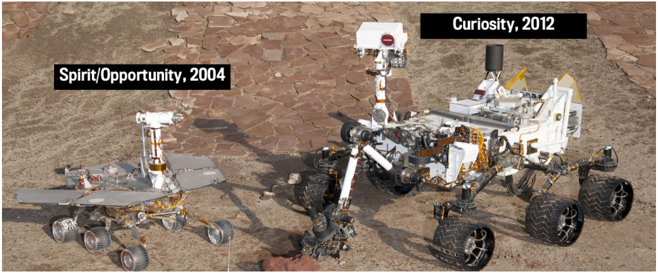
How can learning communication help in RL collaboration?

- Forgoes expensive expert time for protocol planning
- Allows for a decentralised system without an external authority to decide on a communication protocol
- Life-long learning (adaptive tasks, e.g. future proofed robots)

UNIVERSITY OF CAMBRIDGE

How can learning communication help in RL collaboration?

- Forgoes expensive expert time for protocol planning
- Allows for a decentralised system without an external authority to decide on a communication protocol
- Life-long learning (adaptive tasks, e.g. future proofed robots)



From linguistic motivation to a pragmatic view – emergent languages

- Emergent languages
  - Pidgin – a simplified language developed for communication between groups that do not have a common language
  - Creole – a pidgin language nativised by children as their primary language, e.g. *Singlish*

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – computational models

- A computational model for emergent languages should account for
  - polysemy (a word might have different meanings),
  - synonymy (a meaning might have different words),
  - ambiguity (two agents might associate different meanings to the same word),
  - and be open (agents may enter or leave the population, new words might emerge to describe meanings).

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – computational models

- A computational model for emergent languages should account for
  - polysemy (a word might have different meanings),
  - synonymy (a meaning might have different words),
  - ambiguity (two agents might associate different meanings to the same word),
  - and be open (agents may enter or leave the population, new words might emerge to describe meanings).

From linguistic motivation to a pragmatic view – computational models

- A computational model for emergent languages should account for
  - polysemy (a word might have different meanings),
  - synonymy (a meaning might have different words),
  - ambiguity (two agents might associate different meanings to the same word),
  - and be open (agents may enter or leave the population, new words might emerge to describe meanings).

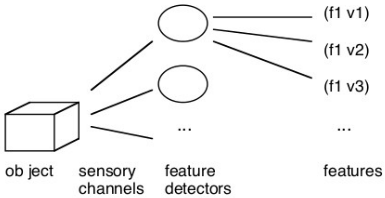
From linguistic motivation to a pragmatic view – computational models

- A computational model for emergent languages should account for
  - polysemy (a word might have different meanings),
  - synonymy (a meaning might have different words),
  - ambiguity (two agents might associate different meanings to the same word),
  - and be open (agents may enter or leave the population, new words might emerge to describe meanings).

UNIVERSITY OF CAMBRIDGE

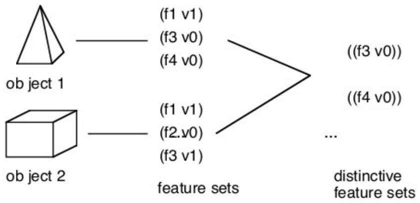
From linguistic motivation to a pragmatic view – computational models

- [Steels, 1996] constructs a model in which words map to features of an object



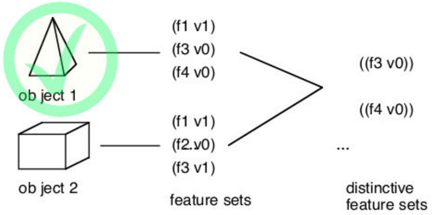
From linguistic motivation to a pragmatic view – computational models

- Agents learn each-other's word-feature mappings by selecting an object and describing one of its distinctive features



From linguistic motivation to a pragmatic view – computational models

- An agent's word-feature mapping is reinforced when both agents use the same word to identify *a* distinctive feature of the object



UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Using RL we can formalise the ideas above
- For example [Goldman et al., 2007] establish a formal framework where agents using different languages learn to coordinate
  - In this framework a state space S describes the world,
  - A_i describes the actions the i'th agent can perform,
  - $F_i(s)$ is the probability that agent i is in state s,
  - Σ_i is the alphabet of messages agent i can communicate,
  - and o_i is an observation of the state for agent i.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Using RL we can formalise the ideas above
- For example [Goldman et al., 2007] establish a formal framework where agents using different languages learn to coordinate
  - In this framework a state space S describes the world,
  - A_i describes the actions the i'th agent can perform,
  - $F_i(s)$ is the probability that agent i is in state s,
  - Σ_i is the alphabet of messages agent i can communicate,
  - and o_i is an observation of the state for agent i.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Using RL we can formalise the ideas above
- For example [Goldman et al., 2007] establish a formal framework where agents using different languages learn to coordinate
  - In this framework a state space S describes the world,
  - A_i describes the actions the i'th agent can perform,
  - $F_i(s)$ is the probability that agent i is in state s,
  - Σ_i is the alphabet of messages agent i can communicate,
  - and o_i is an observation of the state for agent i.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Using RL we can formalise the ideas above
- For example [Goldman et al., 2007] establish a formal framework where agents using different languages learn to coordinate
  - In this framework a state space S describes the world,
  - A_i describes the actions the i'th agent can perform,
  - $F_i(s)$ is the probability that agent i is in state s,
  - Σ_i is the alphabet of messages agent i can communicate,
  - and o_i is an observation of the state for agent i.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Using RL we can formalise the ideas above
- For example [Goldman et al., 2007] establish a formal framework where agents using different languages learn to coordinate
  - In this framework a state space S describes the world,
  - A_i describes the actions the i'th agent can perform,
  - $F_i(s)$ is the probability that agent i is in state s,
  - Σ_i is the alphabet of messages agent i can communicate,
  - and o_i is an observation of the state for agent i.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- We define agent i's policy to be a mapping from sequences (the history) of state-message pairs to actions

$$\delta_i : \Omega^* \times \Sigma^* \to A_i,$$

- and define a secondary mapping from sequences of state-message pairs to messages

$$\delta_i^{\Sigma} : \Omega^* \times \Sigma^* \to \Sigma_i.$$

- A translation τ between languages Σ and Σ' is a distribution over message pairs; each agent holds a distribution $P_{\tau,i}$ over translations between its own language and other agents' languages,
- And meaning is interpreted as "what belief state would cause me to send the message I just received".

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- We define agent i's policy to be a mapping from sequences (the history) of state-message pairs to actions

$$\delta_i : \Omega^* \times \Sigma^* \to A_i,$$

- and define a secondary mapping from sequences of state-message pairs to messages

$$\delta_i^\Sigma : \Omega^* \times \Sigma^* \to \Sigma_i.$$

- A translation τ between languages Σ and Σ' is a distribution over message pairs; each agent holds a distribution $P_{\tau,i}$ over translations between its own language and other agents' languages,
- And meaning is interpreted as "what belief state would cause me to send the message I just received".

From linguistic motivation to a pragmatic view – formal framework

- We define agent i's policy to be a mapping from sequences (the history) of state-message pairs to actions

$$\delta_i : \Omega^* \times \Sigma^* \to A_i,$$

- and define a secondary mapping from sequences of state-message pairs to messages

$$\delta_i^\Sigma : \Omega^* \times \Sigma^* \to \Sigma_i.$$

- A translation τ between languages Σ and Σ' is a distribution over message pairs; each agent holds a distribution $P_{\tau,i}$ over translations between its own language and other agents' languages,
- And meaning is interpreted as "what belief state would cause me to send the message I just received".

From linguistic motivation to a pragmatic view – formal framework

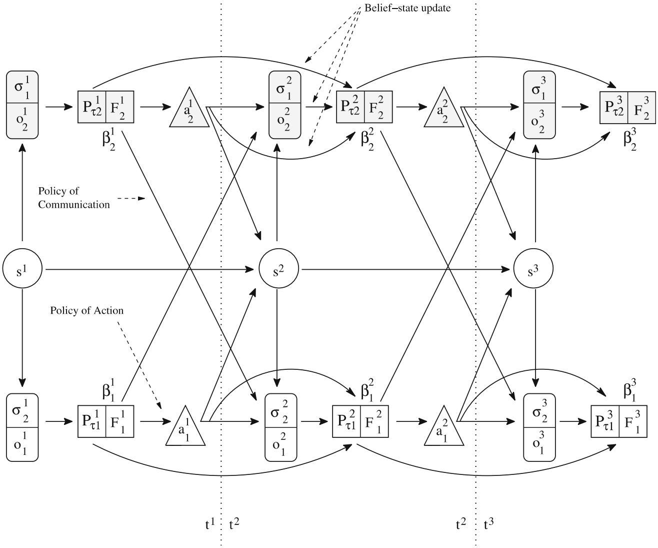
- We define agent i's policy to be a mapping from sequences (the history) of state-message pairs to actions

$$\delta_i : \Omega^* \times \Sigma^* \to A_i,$$

- and define a secondary mapping from sequences of state-message pairs to messages

$$\delta_i^{\Sigma} : \Omega^* \times \Sigma^* \to \Sigma_i.$$

- A translation τ between languages Σ and Σ' is a distribution over message pairs; each agent holds a distribution $P_{\tau,i}$ over translations between its own language and other agents' languages,
- And meaning is interpreted as "what belief state would cause me to send the message I just received".

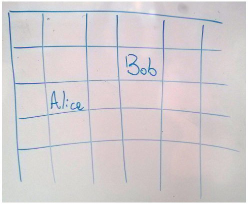


Overview of the framework

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – formal framework

- Several experiments where used to assess the framework.
- For example, two agents work to meet at a point in a gridworld according to a belief over the location of the other.



- Messages describing an agent's location are exchanged and their translations are updated depending on whether the agents meet or not.
- The optimal policies are assumed to be known *before* the agents try to learn how to communicate.

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – a pragmatic view

- Use in robotics
  - A leader robot controlling a follower robot [Yanco and Stein, 1993]
  - Small robots pushing a box towards a source of light [Mataric, 1998]

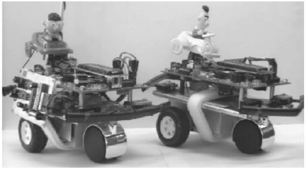


Figure: Leader-follower robots

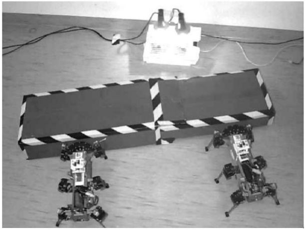
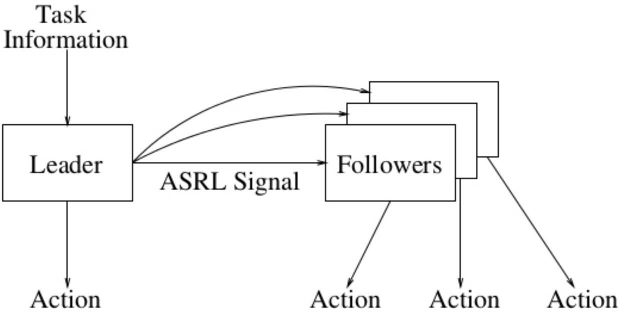


Figure: Box pushing

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – a pragmatic view

- Use in robotics
  - A leader robot controlling a follower robot



Communication diagram

UNIVERSITY OF CAMBRIDGE

From linguistic motivation to a pragmatic view – a pragmatic view

- Use in robotics
  - A leader robot controlling a follower robot

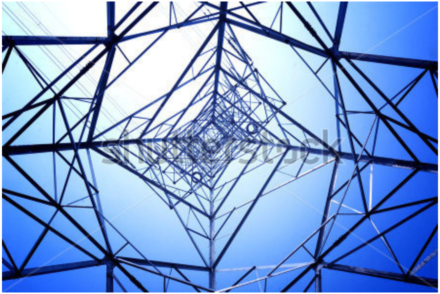
	Appropriate action	Leader's action	Leader's signal	Follower's action	Reinforcement
1.	↑↑	*spin*	low	*spin*	–
2.	◯◯	*spin*	low	*straight*	–
3.	↑↑	*straight*	high	*spin*	–
4.	◯◯	*straight*	high	*straight*	–
5.	◯◯	*spin*	low	*spin*	+
6.	↑↑	*straight*	high	*spin*	–
7.	◯◯	*spin*	low	*spin*	+
8.	◯◯	*spin*	low	*spin*	+
9.	◯◯	*spin*	low	*spin*	+
10.	↑↑	*spin*	low	*spin*	–
11.	↑↑	*straight*	high	*straight*	+
12.	↑↑	*straight*	high	*straight*	+
13.	◯◯	*spin*	low	*spin*	+

Reinforcement regime

Why is learning communication difficult?

What problems exist with learning communication?

- Difficult to specify a framework
  - Many partial frameworks proposed with different approaches
- State space explosion
- Difficult to use for RL collaboration
  - No framework has been shown to improve on independent RL



Why is learning communication difficult?

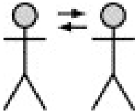
UNIVERSITY OF CAMBRIDGE

What problems exist with learning communication?

- Difficult to specify a framework
  - Many partial frameworks proposed with different approaches
- State space explosion
- Difficult to use for RL collaboration
  - No framework has been shown to improve on independent RL

What problems exist with learning communication?

- Difficult to specify a framework
  - Many partial frameworks proposed with different approaches
- State space explosion
- Difficult to use for RL collaboration
  - No framework has been shown to improve on independent RL



Why is learning communication difficult?

UNIVERSITY OF CAMBRIDGE

What problems exist with learning communication?

- Difficult to specify a framework
  - Many partial frameworks proposed with different approaches
- State space explosion
- Difficult to use for RL collaboration
  - No framework has been shown to improve on independent RL

These problems are not fully answered in current research.

UNIVERSITY OF CAMBRIDGE

Where this might go

- Learning communication based on sparse interactions
  - Reduce state space complexity
- Selecting what to listen to in incoming communication
  - State space selection
- Cyber-warfare – better computer worms?
  - Developing unique communication protocols between cliques of agents
- Online learning of communication
  - Introducing a new agent into a system with existing agents
  - Finding optimal policy with agents ignorant of one another, and then allowing agents to start communicating to improve collaboration

Up, Up and Away

Where this might go

- Learning communication based on sparse interactions
  - Reduce state space complexity
- Selecting what to listen to in incoming communication
  - State space selection
- Cyber-warfare – better computer worms?
  - Developing unique communication protocols between cliques of agents
- Online learning of communication
  - Introducing a new agent into a system with existing agents
  - Finding optimal policy with agents ignorant of one another, and then allowing agents to start communicating to improve collaboration

UNIVERSITY OF CAMBRIDGE

Where this might go

- Learning communication based on sparse interactions
  - Reduce state space complexity
- Selecting what to listen to in incoming communication
  - State space selection
- Cyber-warfare – better computer worms?
  - Developing unique communication protocols between cliques of agents
- Online learning of communication
  - Introducing a new agent into a system with existing agents
  - Finding optimal policy with agents ignorant of one another, and then allowing agents to start communicating to improve collaboration

UNIVERSITY OF CAMBRIDGE

Where this might go

- Learning communication based on sparse interactions
  - Reduce state space complexity
- Selecting what to listen to in incoming communication
  - State space selection
- Cyber-warfare – better computer worms?
  - Developing unique communication protocols between cliques of agents
- Online learning of communication
  - Introducing a new agent into a system with existing agents
  - Finding optimal policy with agents ignorant of one another, and then allowing agents to start communicating to improve collaboration

Where this might go

- Learning communication based on sparse interactions
  - Reduce state space complexity
- Selecting what to listen to in incoming communication
  - State space selection
- Cyber-warfare – better computer worms?
  - Developing unique communication protocols between cliques of agents
- Online learning of communication
  - Introducing a new agent into a system with existing agents
  - Finding optimal policy with agents ignorant of one another, and then allowing agents to start communicating to improve collaboration

Lots to do for future research!

Claus, C. and Boutilier, C. (1998).
The dynamics of reinforcement learning in cooperative multiagent systems.
In *AAAI/IAAI*, pages 746–752.

Goldman, C. V., Allen, M., and Zilberstein, S. (2007).
Learning to communicate in a decentralized environment.
Autonomous Agents and Multi-Agent Systems, 15(1):47–90.

Mataric, M. J. (1998).
Using communication to reduce locality in distributed multiagent learning.
Journal of experimental & theoretical artificial intelligence, 10(3):357–369.

Melo, F. S. and Veloso, M. (2011).
Decentralized mdps with sparse interactions.
Artificial Intelligence, 175(11):1757–1789.

Mundhe, M. and Sen, S. (2000).
Evolving agent societies that avoid social dilemmas.

In *GECCO*, pages 809–816.

Sen, S., Sekaran, M., Hale, J., et al. (1994).
Learning to coordinate without sharing information.
In *AAAI*, pages 426–431.

Steels, L. (1996).
Emergent adaptive lexicons.
From animals to animats, 4:562–567.

Tan, M. (1993).
Multi-agent reinforcement learning: Independent vs. cooperative agents.
In *Proceedings of the tenth international conference on machine learning*, volume 337. Amherst, MA.

Yanco, H. and Stein, L. A. (1993).
An adaptive communication protocol for cooperating mobile robots.
In *Meyer, JA, HL Roitblat, and S. Wilson (1993) From Animals to Animats 2. Proceedings of the Second International Conference on*

Simulation of Adaptive Behavior. The MIT Press, Cambridge Ma, pages 478–485.