Reinforcement Learning for Robotic and Software Agents

Gillian Hayes, gmh@inf

15th October 2008

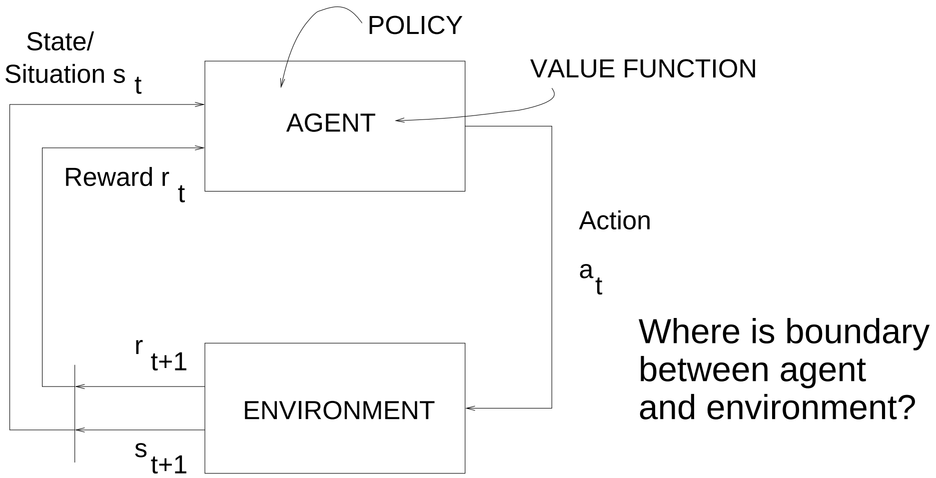
School of informatics

Reinforcement Learning for Robotic and Software Agents

Current work with:

Jay Bradley
Mark Harrison
Matthew Whitaker
Matthijs Snel
Michael Rovatsos
Tom Larkworthy

Reinforcement Learning Reminder



Transition Probability $P^a_{ss'}$: probability of ending up in state s' given that you start in state s and choose action a.

Reward function: if action a chosen in state s and subsequent state reached is s' the expected reward is:

$$R^a_{ss'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\}$$

Learn to act so as to maximise the expected discounted future reward.

We need to **define the reward function** to approximate our expectation of what actions/states will be good.

Markov Decision Process

Topics

- Multi-agent RL with communication
- Perceptual actions in perceptual aliasing
- Game-agent group behaviour: matching the reward function and group structure
- Interactions between (reinforcement) learning and evolution
- Reconfigurable robots
- Feature extraction from EEGs

Multi-agent RL with communication

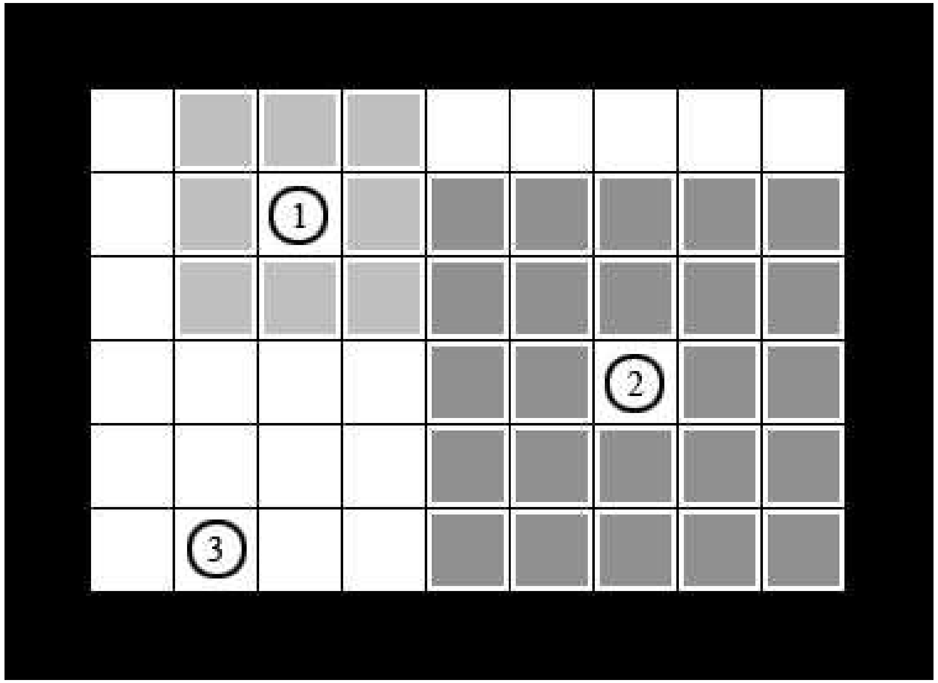
Two agents must move towards each other, goal achieved when adjacent. Movement actions and shout actions. State: observations of squares around agent (grid world). SARSA(λ), true multi-agent RL, **perceptual aliasing**.

Learner shouts, 'homing' agent moves towards it. Optimal policy is for learner to shout all the time.

But actually get **policy segmentation**: if agents are within visual field, move towards other agent, else shout.

Can solve multi-agent perceptually aliased task with communication and without memory.

Communicate actions, Q-tables, states – categorise agents on basis of behaviour; deception?


1
2
3

Perceptual actions in perceptual aliasing

Perceptual aliasing: many states have same state vector, optimal action varies. Need memory to learn optimal policy.

Or use **perceptual actions**, a type of active perception. Instead of moving, look further away from current position. E.g. augment 8-D state vector of squares around current position with the three squares to the northeast – 11-D state vector.

Don't need to search the whole of the 11-D space, just those parts of the space corresponding to aliased 8-D states.

Solves problem for small increase in search, no memory needed.

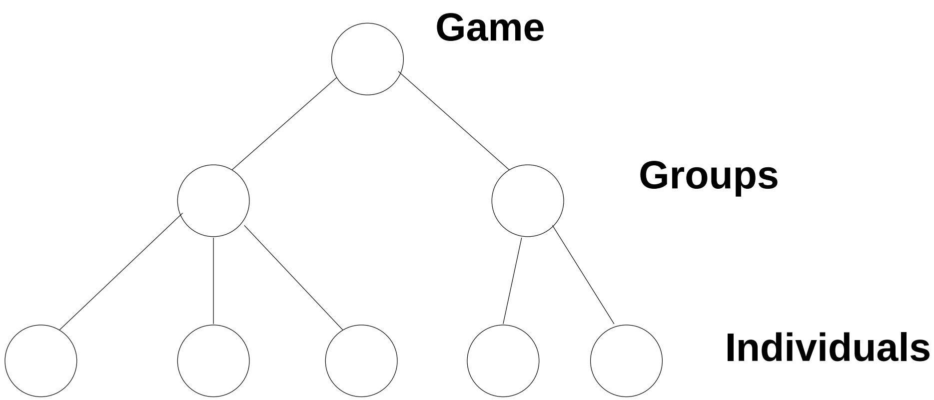
Not guaranteed to converge: if the perceptual states and their corresponding movement states are both aliased at the same time. So pick another set of perceptual states.

Game-agent group behaviour: matching the reward function and group structure

Game agent, e.g. capture the flag. Chooses actions to suit self and the group **and the other – human – player**.

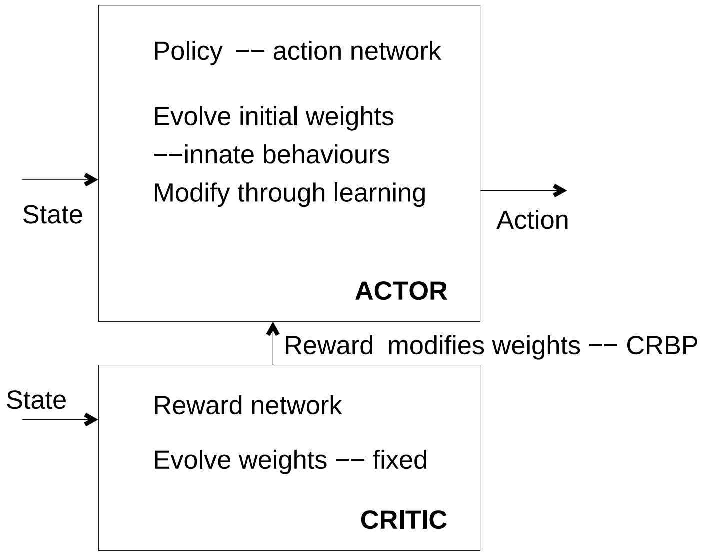
Aim: get a balanced game – more fun for beginner and improver players

Stucture the reward function to match the group structure



Interactions between (reinforcement) learning and evolution

Where does the reward function come from?



CRBP = complementary reinforcement back propagation (a learning method)

Environment: $P^a_{ss'}$ and $R^a_{ss'}$. Make $P^a_{ss'}$ a "natural" consequence of environment, e.g. state = hungry, action = eat food, next state = less hungry – hunger **drive** is satiated. Actions have real consequences for agent – survives or dies.

Why should one action *a priori* be preferred over another?

Evolve the reward function: reward functions that assign the right valency to actions will allow their agents to survive.

Reconfigurable robots

- Make robots out of small actuated units – e.g. rod and spring, blocks and magnets
- Shape is configurable
- Problems: planning how to get from one configuration to another, localisation of units given that the joints are bendy and gravity acts
- Passing through narrow spaces, passing tools around the robot

Feature extraction from EEGs

- Neurofeedback – training someone to control their own EEG
- EEG: measure voltage on scalp, frequency range from about 0 to 40Hz
- Train individual to produce more/less of some frequency ranges at various sites on the scalp, e.g. less 4-8Hz, more 15-18Hz, less $>$ 22Hz in pre-frontal cortex $\rightarrow$ less zoning out, more focussing, less ruminating
- Correlates of brain processes
- Can one detect changes in the EEG that correspond to semi-subjective changes – e.g. alertness