Computational Curiosity

Joseph Modayil
w/ Adam, Patrick, Thomas, Rich
RLAI @ ualberta

April 9, 2010

Computational Curiosity

- Imagine a robot attempting to learn many bits of knowledge about the world.
  - How long does it take to walk over this hill? How long does it take to walk around it?
  - Will I see a light if I drive forward for 5 timesteps?
  - Will I run out of power in five minutes?
  - Will I see a T-junction if I follow this wall?
  - How soon can I get to the docking station?
  - How do I stop as quickly as possible?
- What *behaviour* allows an agent to efficiently learn answers to many questions simultaneously?
- Proposal: Use curiosity-driven behaviour that is rewarded for progress in learning answers.

Standard RL machinery

- S: a set of states
- A: a set of actions
- r_t: a reward at each time step
- $Q(s, a) \rightarrow \Re$: a state-action value function
- $\pi(a|s) \rightarrow [0, 1]$: a policy

Demon knowledge

We consider a demon $h =< q, o, v >$ learning an answer to a single option-conditional question.

- An option $o =< I, \pi, \beta >$ is an abstraction of temporally extended actions.
  - $I \subset S$ is the set of interest.
  - π is the target policy.
  - $\beta : S \rightarrow [0, 1]$ is the termination function.
- The question $q =< r, z >$ is about the expected return from following the option o to termination.
  - $Q^h(s_t, a_t) = E[r_{t+1} + \ldots + r_T + z_T | s = s_t, a = a_t]$
  - r_t is the transient reward at time t.
  - z_t is the outcome reward at option termination at T.
- The answer is learned as a linear function of the state-action feature vector
  - $\tilde{Q}^h(s, a) = v^T x(s, a)$

Questions about goals

Many interesting questions are about some goal set $G \subset S$ where the option preferentially terminates.

- Consider a termination probability $\beta = \epsilon + 1_G$
- Then the return from $z = 1_G, r = 0$ is the probability of terminating in the goal.
- We could measure time to termination ($r = -1$).
- We could have auxiliary constraints $\beta = \epsilon + 1_G + 1_{Unsafe}$.

Questions about induced policies

Questions may be about an (unknown) best policy for performing a task.

- How soon can I get home from here?

In this case, a target policy is induced from the current value function estimate.

GQ Learning

We can use the recently developed GQ algorithm to learn the answers through off-policy behaviour. Benefits include:

- Off-policy learning with arbitrary function approximation
- Stability proofs (for fixed behaviour)
- Online algorithm
- Complexity increases by a small factor.

One agent, many demons

We have a single agent with a horde of demons $H = \{h \mid h =< q, o, v >\}$ in its head. How should the agent behave to learn the answers to all the demon questions?

- We define an internal *curiosity* reward from the sum of learning progress across the demons.

$$r^c_{t+1} = \frac{1}{|H|} \sum_{h \in H} ||v^h_{t+1} - v^h_t||^2$$

- We can measure predictive performance using the difference between the estimate and true answers.

$$J_P(H) = \frac{1}{|H|} \sum_{h \in H} ||Q^h - \tilde{Q}^h||^2_{De}$$

Here, De is the state-action distribution of an evaluation policy e.

Possible Behaviours

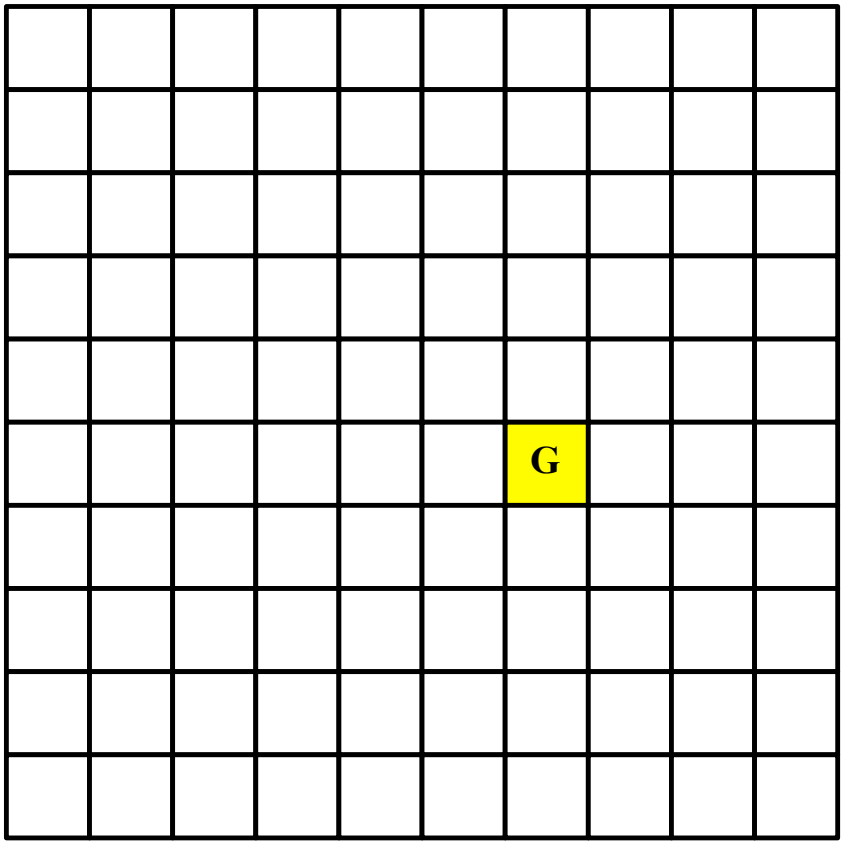
- Random
- Evaluation
- Curiosity
- Evaluation + Random
- Evaluation + Target
- Evaluation + Curiosity

Switching in mixed behaviours can be performed at each timestep, or in phases. Initialize target behaviours with zero, and curiosity behaviour optimistically.

Domains

- Gridworld with weak function approximation
  - 10×10 states, 5 actions,
  - 200 bit feature vector, 10 bits on at random for each state + 1 bias bit.
  - ϵ-greedy ($\epsilon = .1$)
- Problems
  - Single goal
  - Multiple goals
  - Hierarchical goals

Results: Single Goal

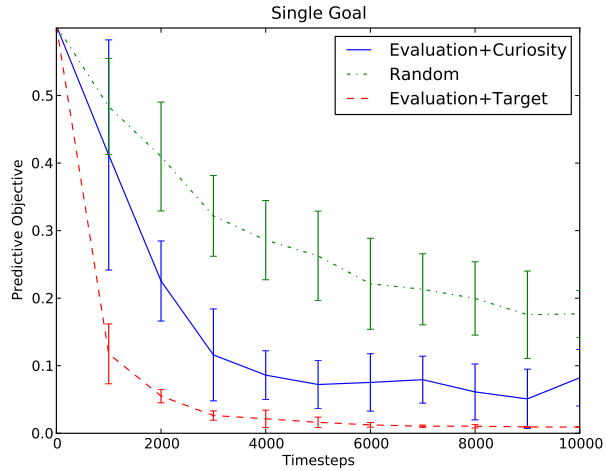


evaluation policy=Random

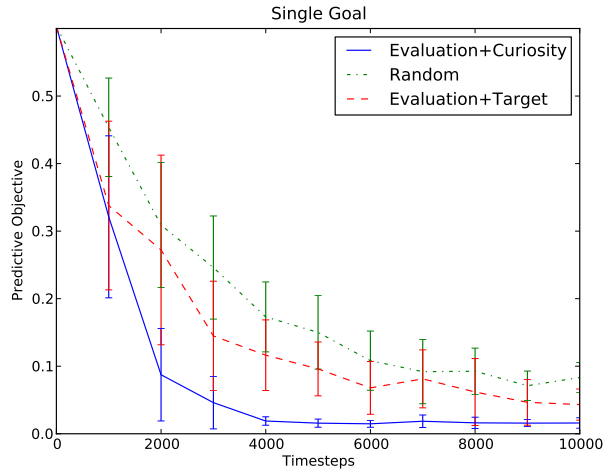
$r = 0$; $z = 1_G$,

$\beta = 1/20 + (1 - 1/20)1_G$

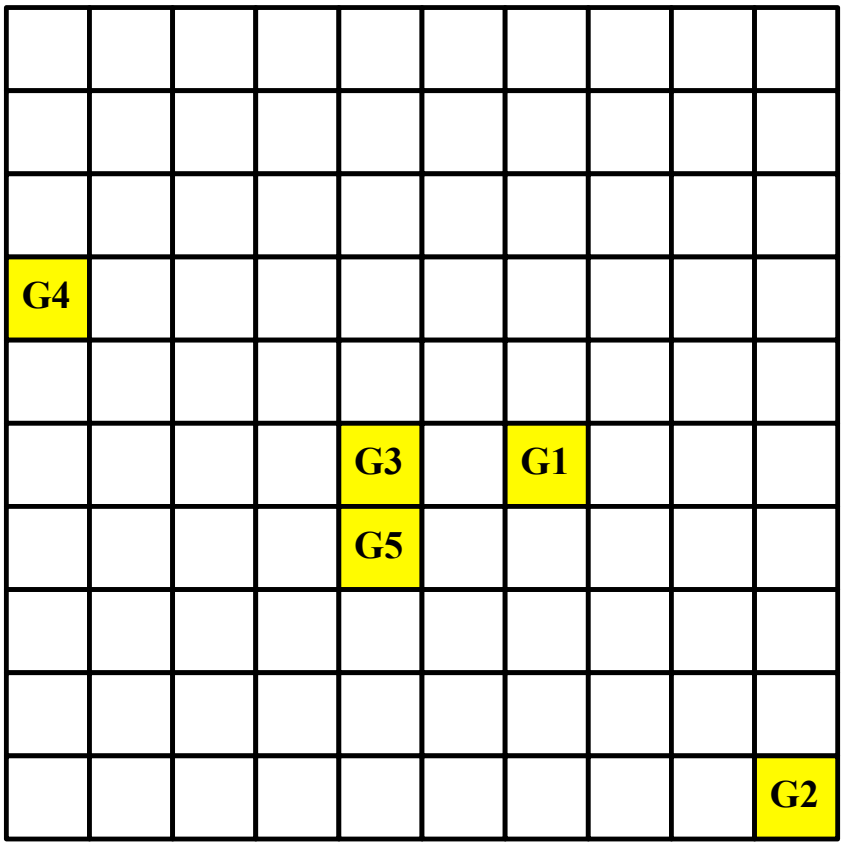
Fixed Policies



Induced Policy



Results: Multiple Goals

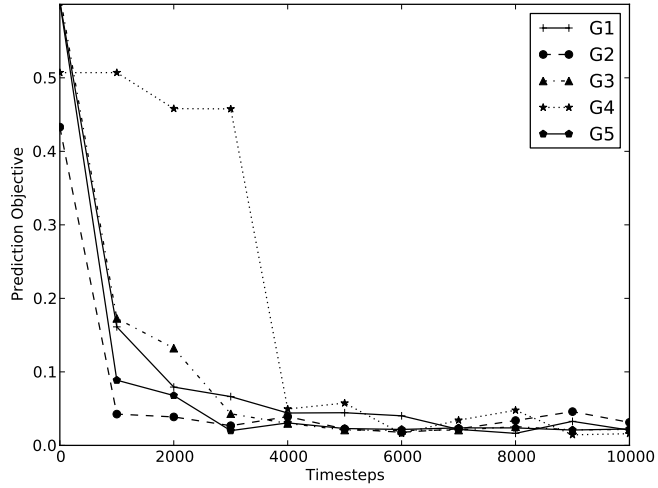


evaluation policy=Random

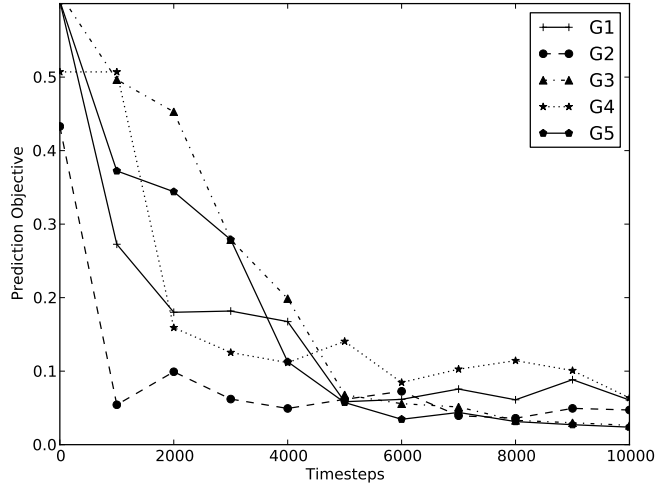
$r = 0$; $z = 1_G$,

$\beta = 1/20 + (1 - 1/20)1_G$

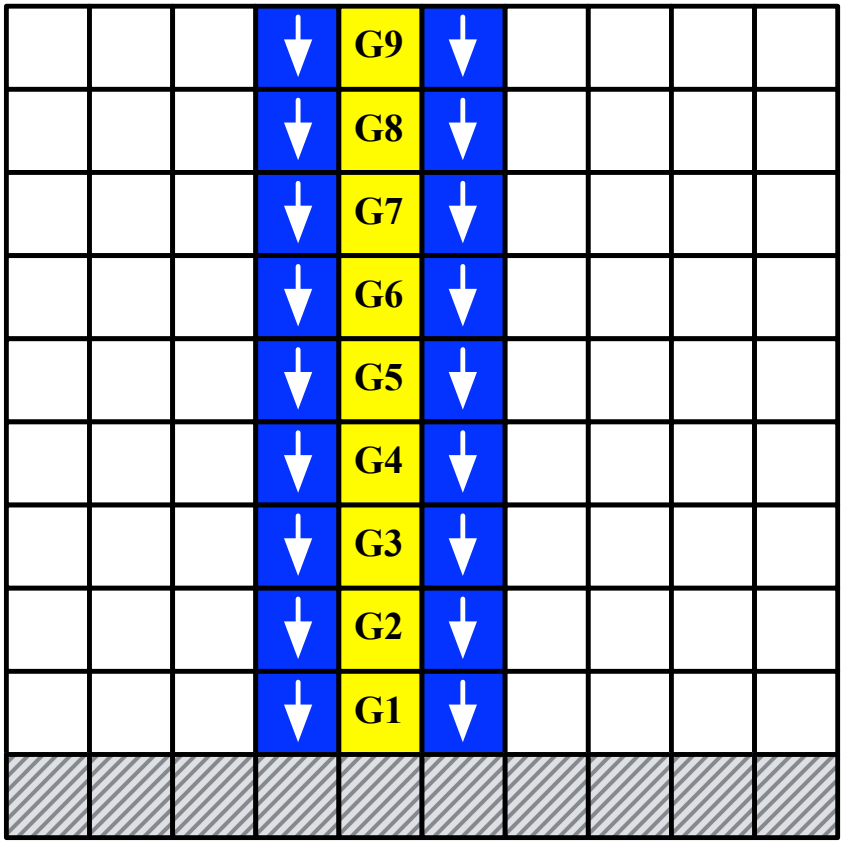
Curiosity Behaviour



Target Following



Results: Hierarchical Goals

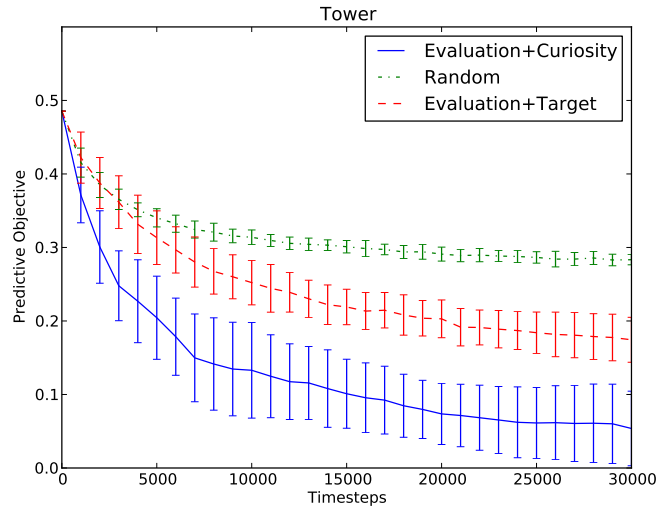


evaluation policy =Random motion in shaded region

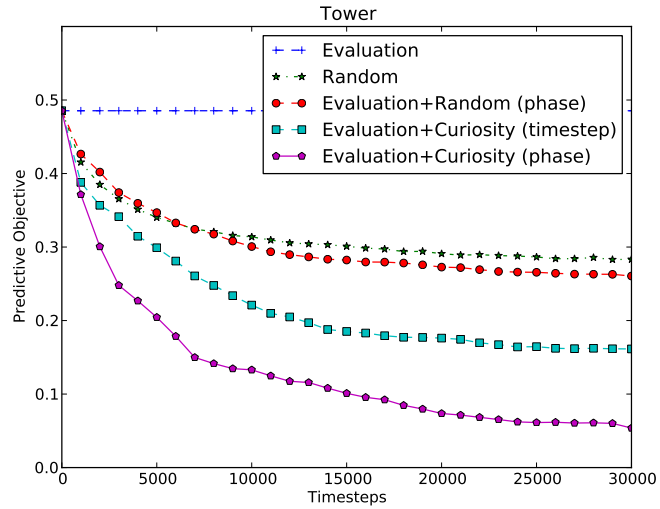
$r = 0$; $z = 1_G$,

$\beta = 1/20 + (1 - 1/20)1_G$

Curiosity



Alternatives



On going: Robot Goals

Consider the critterbot learning to answer many questions.

1. How do I minimize IR reflectance sensor 2 ?
2. How do I move so reflectance sensor 2 is the maximum of the distance sensors?
3. How do I maximize reflectance sensor 2 ?
4. How do I move wheel 1 forward while maximizing reflectance sensor 2? (wall-following)

Evaluate by measuring performance and comparing to predictions. (Current status video)

Related Work

- Curiosity proposed as chasing novelty (Schmidhuber 1991a) and then as chasing learning progress (Schmidhuber 1991b)
- Curiosity can drive an agent to learn about hierarchical options (Singh, Barto, Chentanez 2005) with a planning based approach.
- Curiosity can drive an actual robot (single step prediction) (Oudeyer et al. 2007).

We are studying curiosity-driven exploration without planning.

Conclusions

- GQ is a stable, online, off-policy algorithm. We can use it to learn to answer many questions simultaneously
- An agent's ability to pose questions and answer them is limited by its representations.
- Curiosity is a useful exploration behaviour.