Using Active Relocation to Aid Reinforcement Learning

Lilyana Mihalkova and **Raymond Mooney**
University of Texas, Department of Computer Sciences,
1 University Station, C0500,
Austin, TX 78712
{lilyanam,mooney}@cs.utexas.edu

AAAI 2006

Relocate!

Motivation

Sample inefficiency

- Just like a person learning to play game, some stages of the game are easy to pass.
- While trying new approaches, the player makes a silly mistake that would inevitably end the game.

Methods

- When to Relocate

- Where to Relocate

Methods – When to Relocate

- Agent is "In Trouble"

  - At this time, players often hope to reset as soon as possible, rather than try to find a way to save all this.

$$\rho = \rho + \nu(Q(s,a) - Q(s',a^*))$$

  - With the state s, the agent takes an action a and receives the next state s'. Even the best action a^* is taken, the value of Q will decrease sharply, it can be say the Agent is "In Trouble".

Methods – When to Relocate

- Agent is "Bored"
  - It means that for some state-action pairs, the agent can only get a small amount of updates. In other words, the agent cannot learn from these trajectories.

$$\rho = \begin{cases} \frac{1}{2} e^{\frac{\delta \cdot \phi}{\Delta Q_{min}}} & \text{if} \quad \delta \in [-\Delta Q_{min}, 0); \\ \frac{1}{2} e^{\frac{-\delta \cdot \phi}{\Delta Q_{max}}} & \text{if} \quad \delta \in [0, \Delta Q_{max}]. \end{cases}$$

  - Where ϕ is a hyperparameter, ΔQ_{min} and ΔQ_{max} are the largest decrease and increase, respectively.

Methods – Where to Relocate

- Two requirements:
  - It should be likely to be encountered while following an optimal policy.
  - And it should be one in which the agent is uncertain about the best action.

- Two solutions:
  - 1. Since the optimal policy is unknown, the agent cannot be sure whether a state is relevant. However, as the agent accumulates more experience, its policy is likely to come closer to optimal.
  - 2. State with the most uncertain best action of agent.

Methods – Where to Relocate

The agent's uncertainty is measured as the width of the confidence interval of the difference between the means of the two best Q-values in a given state s.

$$uncert(s) = 2 \cdot t_{\alpha/2,m+n-2} \cdot v_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

In this formula m and n are the number of times respectively each of the two best actions has been attempted. $t_{\alpha/2,m+n-2}$ is the critical value for the t distribution at confidence level α with $m + n - 2$ degrees of freedom. v_p is given by:

$$v_p^2 = \frac{(m-1) \cdot v_1^2 + (n-1) \cdot v_2^2}{m+n-2}$$

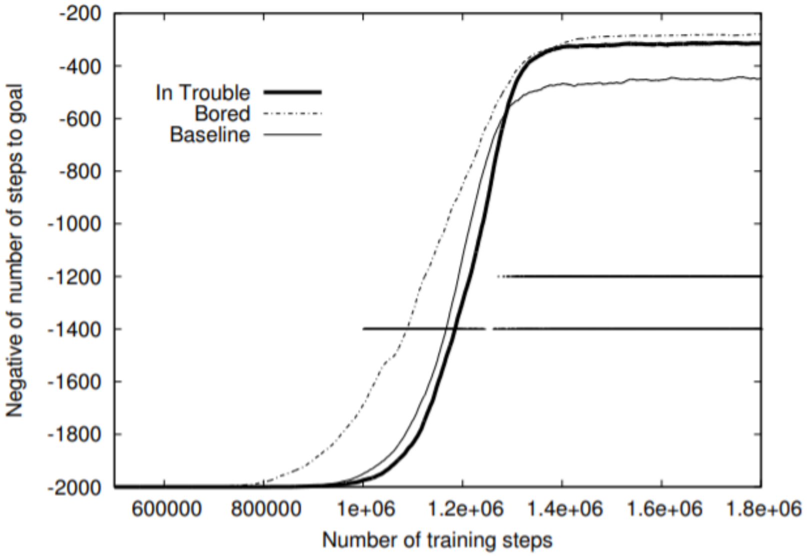
where v_1^2 and v_2^2 are the sample variances given by:

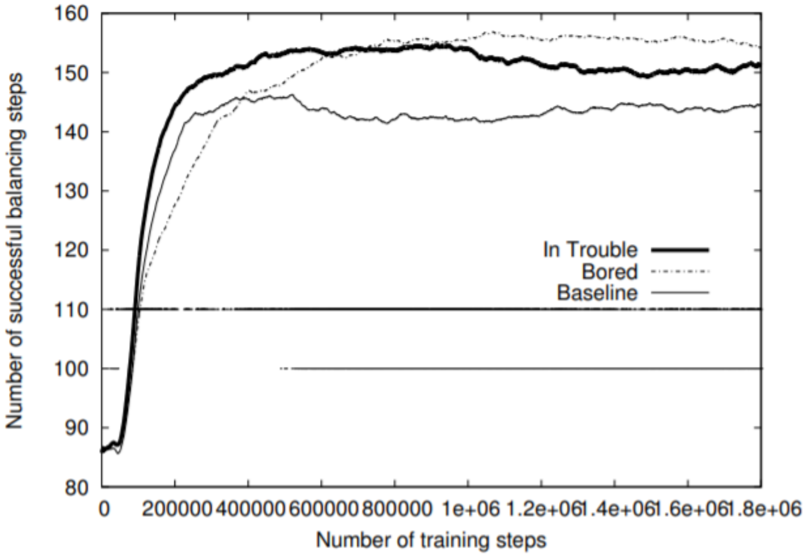
$$v_1^2 = \frac{\sum_i q_i^2 - (\sum_i q_i)^2/m}{m-1}$$

For a particular (s, a_1) pair, the q_i's are obtained by "sampling" $Q(s, a_1)$ each time the action a_1 is taken in state s.

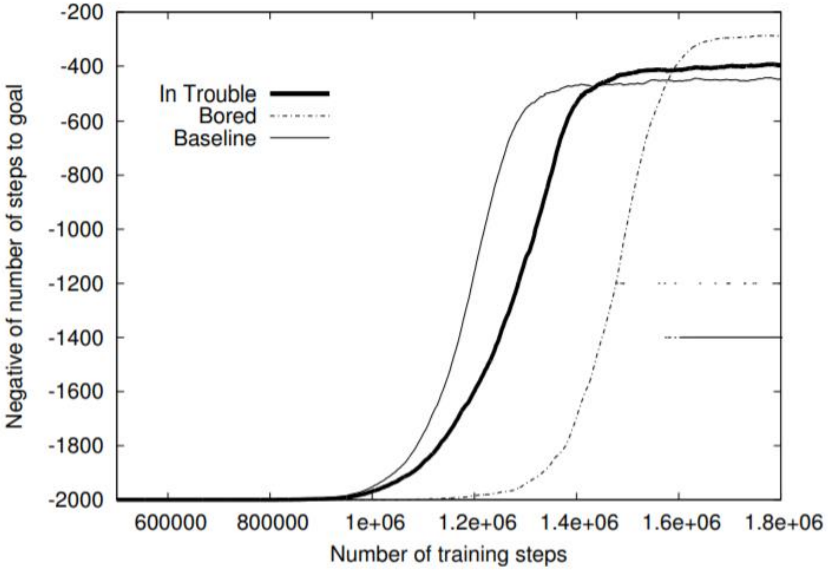
Experiments

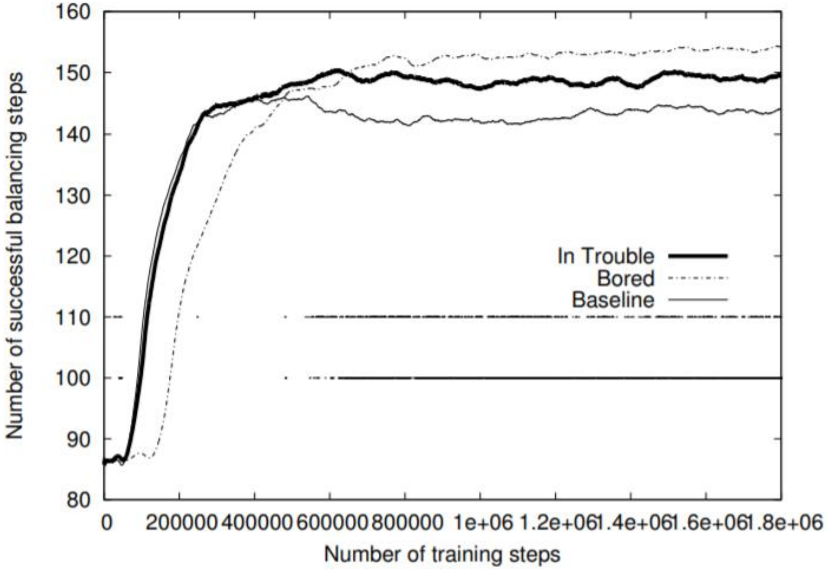
Higher probability to Relocate





Lower probability to Relocate





THANKS