A Comparison of Reinforcement Learning Methods for Automatic Guided Vehicle Scheduling

DoKyeong Ok*
Department of Computer Science
Oregon State University
Corvallis, OR 97331
okd@research.cs.orst.edu

Automatic Guided Vehicles or AGVs are increasingly being used in manufacturing plants for transportation tasks. Optimal scheduling of AGVs is a difficult problem. A learning AGV is very attractive in a manufacturing plant since it is hard to manually optimize the scheduling algorithm to each new situation.

In this paper we compare four reinforcement learning methods for scheduling AGVs. Q-learning[Watkins and Dayan 92] and R-learning[Schwartz 93] do not use action models. Q-learning optimizes the discounted total reward, while R-learning optimizes the average undiscounted reward per step. ARTDP[Barto et al. to appear] is a discounted method that uses action models. H-learning[Tadepalli and Ok 94] is an undiscounted version of ARTDP based on an algorithm of Jalali and Ferguson[Jalali and Ferguson 89].

In our domain(see Figure 1), there are two queues generating jobs, an AGV, a moving obstacle and two lanes. Queue 1 generates jobs for lane 2 half the time and Queue 2 always generates lane 1 jobs. The task of AGV is to move jobs from the queues to their destination lanes while avoiding collisions with the obstacle, which randomly moves up and down. There are a total of 540 states. At any time an AGV may do nothing, load, unload, or move up, down, left or right. The goal is to maximize the average reward per step.

An experiment compared Q-learning, R-learning, ARTDP and H-learning in our AGV domain. The reward is -5 when the AGV collides with the obstacle, +5 when it unloads a job to lane 1, and +1 when it unloads a job to lane 2. Figure 1 shows the medians of average reward per step over 30 trials evaluated separately after turning off learning at various stages. To enable exploration, 50% of the time a randomly chosen action was excuted during learning. The parameters of ARTDP, Q-learning, and R-learning were tuned to this domain by trial and error. ARTDP with discount factor γ=0.9 and Q-learning could not converge to the optimal policy even after 2 million steps. Even though R-learning and ARTDP with high γ converged to the optimal policy, they converged slower than H-learning. The results show clearly that H-learning converges to the optimal policy fastest without any parameter tuning, while the other three methods are very sensitive to the parameters.

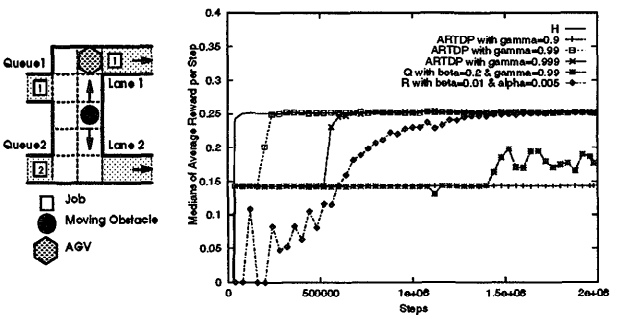

Figure 1: An AGV domain(left); average reward per step for the four learning methods(right)

The future research will explore extensions to H-learning that scale for larger state spaces.

References

Barto, A. G., Bradtke, S. J., and Singh, S. P. To appear. Learning to Act using Real-Time Dynamic Programming. *Artificial Intellignece.*

Jalali, A. and Ferguson, M. 1989. Computationally Efficient Adaptive Control Algorithms for Markov Chains. In IEEE proceedings of the 28th Conference on Decision and Control, Tampa, FL.

Schwartz, A. 1993. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In proceedings of the Tenth International Machine Learning Conference, 298–305. San Mateo, CA.:Morgan Kaufmann.

Tadepalli, P. and Ok, D. 1994. H-learning: A Reinforcement Learning Method to Optimize Undiscounted Average Reward, Technical Report, 94-30-1. Dept. of Computer Science, Oregon State Univ.

Watkins, C. J. C. II. 1989. Learning from Delayed Rewards. Ph.D. Thesis, Cambridge univ., Cambridge, England.

*This research was supported by the National Science Foundation under grant number IRI:9111231. I thank my advisor Prasad Tadepalli for his help and guidance.