Game Day #1 - Post-game Report

Kiatten Mittons - Nathan DeMaria, Rees Klintworth, Derek Nordgren

February 12, 2015

Reflections

For Round 2, we chose our beta to be 0.70 and our alpha to be 0.50 - sqrt(t/(3m)), where t is the time elapsed since the beginning of the round in minutes and m is the duration of the round in minutes (we estimated this to be 20). With this selection of alpha, our algorithm converged to zero with roughly five minutes left in the round. At this time, due to an alpha of zero, future state-action pairs would no longer update our Q-table. We put away our application and Q-table (as we had learned an optimal path) and were able to execute actions yielding optimal rewards at a high rate without sacrificing a point deduction for not updating our Q-table. Having a larger beta than Round 1 caused our algorithm to be more forward-looking and seek maximum future rewards.

One trial we faced was keeping our alpha (expressed as a function of time) in sync with our application's alpha (expressed as a constant). In order to do so, a member of our team had to calculate alpha every other minute or so and we had to manually update the application's algorithm. Were we to play again, we would have the application update alpha automatically based on the system time.

We believe that our strategy was well-conceived and well-executed. It made good use of an exploration period in the first round, but then gave ample time to exploit the results of the exploration period. However, we believe that we could have had our alpha set to converge earlier in the second round. It was important to create a well populated and accurate Q-table, but by the second round we had explored the majority of paths at least once, and could determine the best patterns to follow. This is due in part to the fact that we did not anticipate that the set of state-action pairs would be so small (6x6). Therefore, we were able to amply explore the problem and determine a reliable Q-table faster than we had anticipated. A faster alpha convergence would have enabled us to exploit at a high rate for a longer period of time.

Statistics

Amount of rewards:	$257,181
Sales:	$1,000
Purchases:	-$0
Total:	$258,181

Amount of rewards after Round 1: $35,045.70
Total actions taken: 224 (second-most in the game day)
Round 2 actions taken: 173
Rank (by rewards): 1

We were able to achieve this relatively high number of actions taken because of our alpha function. When alpha is zero, the Q-table is no longer updated. Our confidence in our Q-table values and our knowledge of a high-reward path at the mid-session break allowed us to pick a function for alpha that approached zero quickly. This meant that we could then stop entering values into the application, and as a result our action selection time dropped significantly. As a result of this increase in speed, we were able to overtake the first place team in total rewards before the end of Round 2.

Conclusion

Through this Game Day experience, our team learned first hand about the trade-offs involved with a learning algorithm such as Q-learning. Although we attempted to balance exploration and exploitation, by the end of the second round we had concerned ourselves fully with obtaining a high overall reward. We also gained hands-on experience in determining when and how to change alpha and beta values to tune a Q-learning algorithm based on the state-action environment it is running in. Additionally, we learned how to reflect a desired strategy through the algorithm. In the end, we saw first-hand how a simple algorithm like Q-learning can lead to powerful results.