Daniel Fernández
932-063-474
ME538
3 November 2014
Homework #2

The assignment calls for the modelling and simulation of W. B. Arthur's El Farol bar problem, a congestion game used as an agent learning example. The simulation assumes there is a sole bar in town and a collection of agents which must choose which night is best to attend. This is solved by creating a system reward for achieving the optimal amounts of agents at the bar per night. This optimal value is preset and arbitrary assuming that a positive experience is achieved when attendance is neither too high nor too low. The analysis will focus on three agent reward structures: a simple "local" reward, a difference rewards, and a system reward modeled by G as shown:

$$G(z) = \sum_{k=1}^{K} x_k(z) * e^{\frac{-x_k(z)}{b}}$$

Where G is the anticipated reward for a specific night, x is the total attendance on a given night, k, K is the total number of nights, and b is the aforementioned "optimal" value of attendees per night. Each agent will keep a K-long vector of G estimates and each iteration, or week, the agents will choose the best action based on prior experience.

Part 1 – "Local" Reward Derivation

The Local reward is plotted after each agent selects a night. The agents select a bar based on a "move" function which uses a greedy-epsilon method of maximizing the reward. Epsilon is set to have a 1:100 chance of moving randomly. All agents also move randomly during the first week of a given episode. A K-long vector "patrons" is populated with the amount of agents at the bar per night. This vector is then passed through a "rewards" function which calculates how each agent performed based on the optimal value "b". The local reward is then added to the system reward. This reward structure is good at spreading around the agents, but it does not converge on a scenario where the maximum number of optimal nights is achieved during a week. Therefore, the factoredness is lower than other methods.

Part 2 – Difference Reward Derivation

The Difference Reward focuses on two counterfactuals, one where each agent is removed and then either added randomly elsewhere, or never added back at all. The counterfactual implemented is one where the agent action is subtracted and the reward is recalculated. The agent then is conditionally reinserted should the reward be greater. This provides for improved factoredness as the agents essentially interpret two potential rewards for each one that the other methods perform per week. There is also a higher likelihood that more "optimal" nights where attendance is equal to the value for "b". The learnability, on the other hand, is decreased as the learning curve is steeper and takes more weeks to develop.

Part 3 – Simulation Results

For the first simulation, the number of available nights, K, is set to 6 and the amount of available agents is set to 30. Figures 1 and 2 show sampled histograms for attendance in week 750 of 1000, one for the System Reward, one for the Difference reward respectively. The system reward shows a noticeable distribution among the nights of the week. Of interest is day 2, where 0 agents are in attendance. By way of a visualization, this anomaly became apparent; every few trials, the agents would avoid a particular night over a series of several hundred weeks before eventually converging upon it. It happened too sporadically to fully understand. Figure 2 illustrates the improved factoredness of the Difference method, where “optimal” nights happen more frequently. Here, 4 of 6 days have the preferred attendance of 4 patrons.

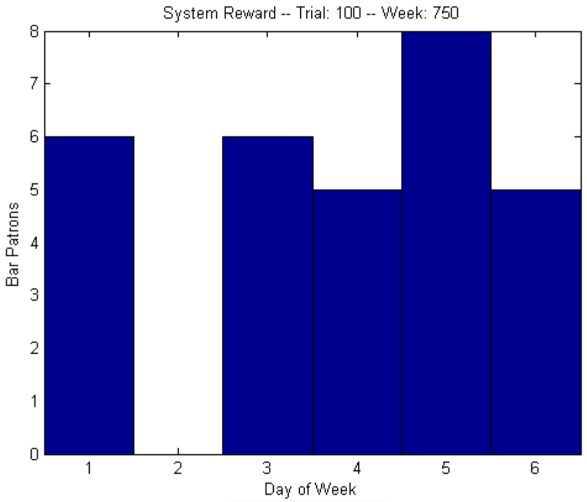

Figure 1

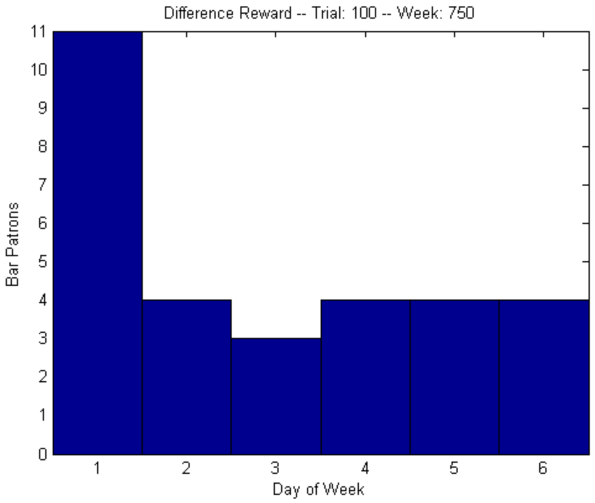

Figure 2

Figure 3 below shows the system performance for all 3 cases. The local reward was the poorest performer while the difference reward performed the best after several dozen weeks. Of interest are the less steep learning curves for the System and Local rewards. This is due to the fact that their earlier moves are influenced by random action rather than a reward structure, which leads to a more even distribution and quicker learnability, as the reward arrays are more rapidly populated. Of another interest is the depreciating System Reward. This was an issue only in this first simulation.

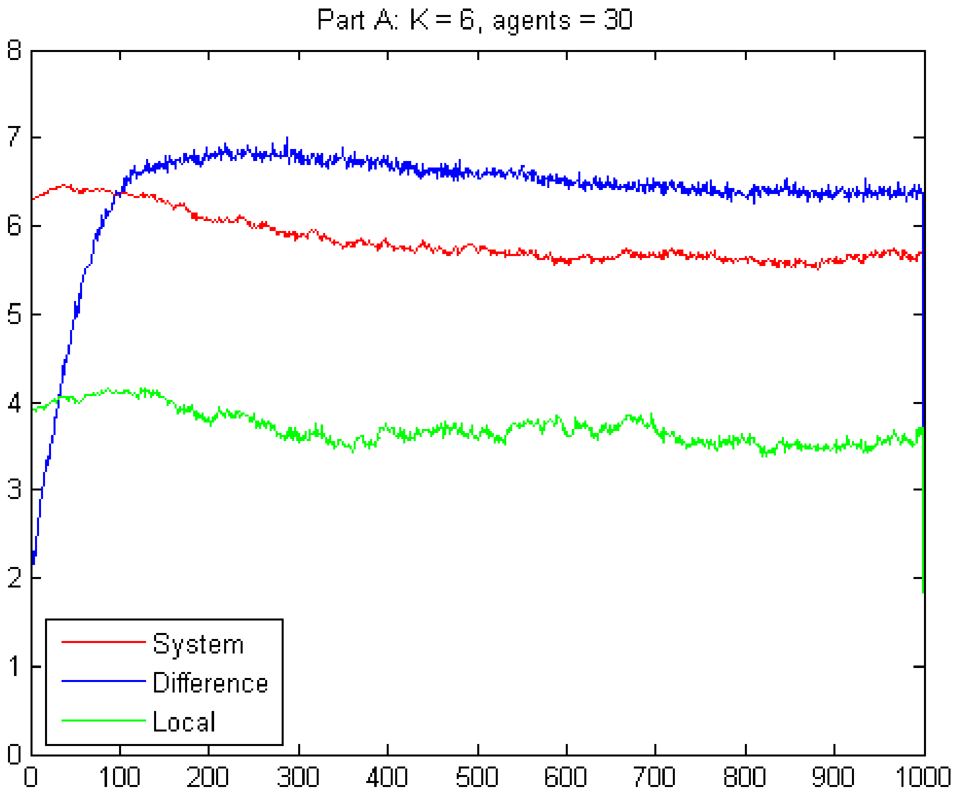

Figure 3

In the second simulation, the number of available nights, K, is set to 5 and the amount of available agents is set to 50. Similar to the figures above, figures 4 and 5 show sampled histograms for attendance in week 750 of 1000, one for the System Reward, one for the Difference reward respectively. Here, the system reward shows a less balanced distribution among the nights of the week than shown in Figure 2. Figure 5 again illustrates the improved factoredness of the Difference method, as “optimal” nights still happen more frequently with 3 of 5 days having the preferred attendance of 4 patrons.

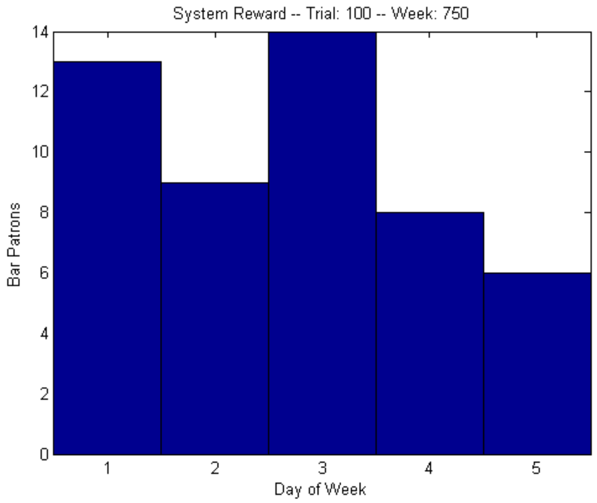

Figure 4

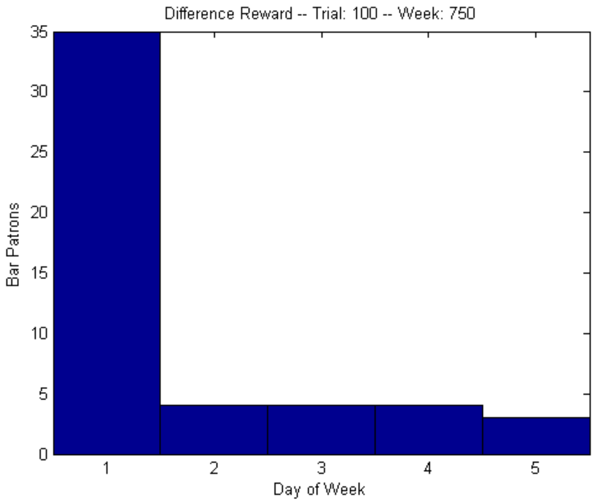

Figure 5

Figure 6 below shows the system performance for all 3 cases. Again, the local reward was the poorest performer while the difference reward remained the best. The learning curves for the System and Local remained while the curve for the Difference Reward remained steeper. The overall reward values are lower, as expected since it is harder to reach an optimal state with more agents and fewer available nights. The depreciating System Reward is missing from this simulation however, and more investigation into this phenomenon is suggested.

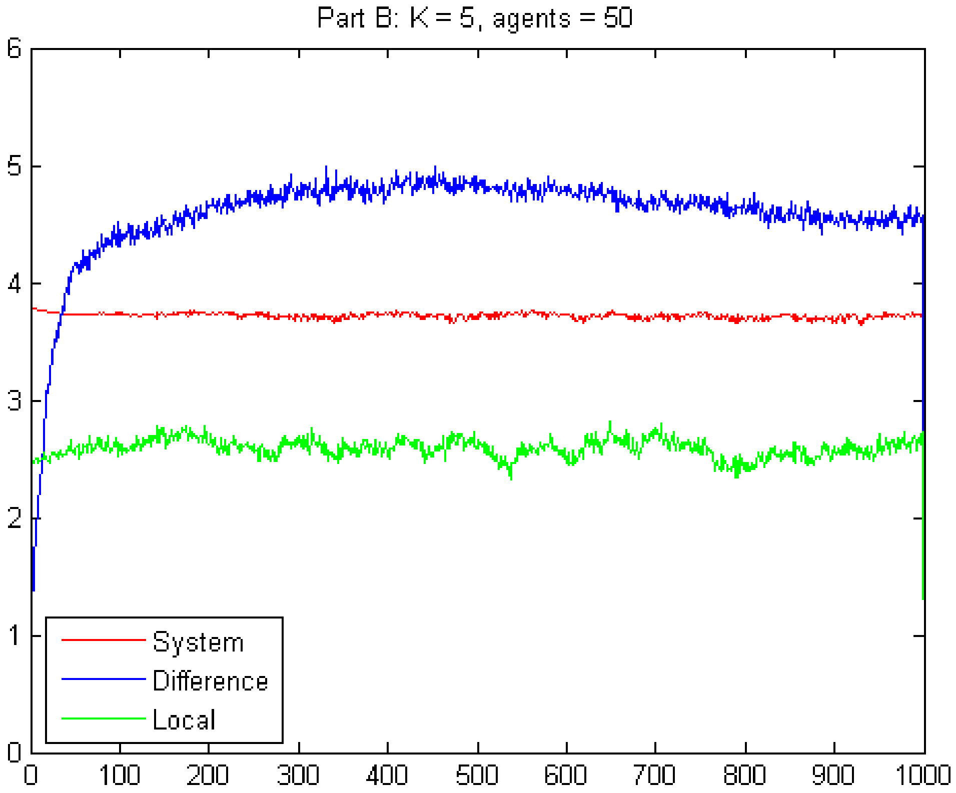

Figure 6