### CSCE475H Multiagent Systems

# Handout 16: Game Day 1 Learning Day Analysis

### February 16, 2015

## State Transition Map and Rewards

There were six states (S1-S6) and six actions (a1-a6). Each team started with S1. Each team was capable of performing all six actions. Table 1 shows the rewards for transitioning into each state and, its average and standard deviation, based on a Gaussian distribution.

| | Average | Std. Dev. |
|---|---|---|
| **S1** | $0 | $10 |
| **S2** | $100 | $10 |
| **S3** | $1500 | $10 |
| **S4** | $500 | $10 |
| **S5** | $5000 | $10 |
| **S6** | $1000 | $10 |

**Table 1.** Rewards, average and standard deviation values, Gaussian distribution.

Table 2 shows the probabilistic transition map for each state-action pair. Looking at both Tables, if one aimed to obtain the highest reward for a state (i.e., S5 @ $5000), then starting for S1, one would probably have to go with a4 to transition into S4 (with a high probability @ .7), and then go with a5 to transition into S5 (with a high probability @ .9). And then to get back to S1, one could perform an action of a6, if so desired. This sequence of a4-a5-a6, when repeated, should allow an agent to reach S5 with a relatively high probability (= .7 x .9 x .5 = .315), and a relatively high reward (= $500 + $5000 + $0 = $5500). With enough exploration, an agent should be able to discover this sequence.

| | a1 | a2 | a3 | a4 | A5 | A6 |
|---|---|---|---|---|---|---|
| **S1** | → S1 (.50)<br>→ S2 (.30)<br>→ S3 (.15)<br>→ S4 (.05) | → S2 (.80)<br>→ S3 (.20) | → S2 (.10)<br>→ S3 (.90) | → S1 (.05)<br>→ S2 (.25)<br>→ S4 (.70) | NA | NA |
| **S2** | → S1 (.70)<br>→ S2 (.15)<br>→ S4 (.15) | → S1 (.55)<br>→ S3 (.35)<br>→ S4 (.10) | NA | NA | NA | → S5 (.50)<br>→ S6 (.50) |
| **S3** | NA | NA | → S1 (.70)<br>→ S2 (.20)<br>→ S4 (.10) | → S1 (.60)<br>→ S3 (.30)<br>→ S4 (.10) | NA | NA |
| **S4** | → S1 (.60)<br>→ S2 (.20)<br>→ S3 (.20) | → S1 (.65)<br>→ S2 (.14)<br>→ S3 (.20)<br>→ S4 (.01) | → S1 (.98)<br>→ S2 (.02) | NA | → S5 (.90)<br>→ S6 (.10) | NA |
| **S5** | NA | NA | NA | NA | NA | → S1 (.50)<br>→ S2 (.30)<br>→ S3 (.15)<br>→ S4 (.05) |
| **S6** | → S1 (1.0) | → S2 (1.0) | → S3 (1.0) | → S4 (1.0) | NA | NA |

**Table 2.** State transitions by actions. NA for a state-action cell means the action is not applicable for the state.

Because of the limited time on Game Day, we did not expect teams to obtain accurate Q-values. However, teams should be able to obtain fairly accurate ordering of their Q-values. The ordering of the best state-action pairs (Q(s,a)) is as follows:

Group 1: (S5,a6)
Group 2: (S4,a5); (S5,a1); (S5,a2); (S5,a3); (s5,a4); (s5,a5)
Group 3: (S2,a6); (S6,a4)
Group 4: (S3,a3); (S3,a4); (S4,a4); (S6,a2); (S6,a3)
Group 5: (S1,a4); (S4;a1); (S4,a2); (S4,a6); (S6,a1); (S6,a5); (S6;a6)

For the above, we also define a function called Group_true(s,a) that returns the group ID of a state-action pair. So, for example, Group_true(S5,a6) is 1; Group_true(S4,a6) is 2; Group_true(S5,a1) is 2; and so on.

## Team Statistics

Tables 3 and 4 show the ordering of the teams after Round 1 and Round 2, respectively.

To compute the accuracy of a Q-table, we use the grouping shown earlier. We consider only the top 18 state/action pairs in each team's Q-table (where 18 is half of the 36 possible values). (Important Note: the last group actually only has 4 elements (not 7) when we limit ourselves to only looking at the top 18 for each group. We however put 7 state/action pairs in Group 5 to be fair to teams since they are all pretty equivalent in that group, and using only 4 would mean teams wouldn't get credit if they had the other 3 (equivalent) pairs, instead.)

First, we sort each team's Q-values.

And second, for each state-action pair on the sorted list, we assign Group_found(s,a) using the 1-6-2-5-7 grouping strategy. So, take Leen Dream's Round 1 ordering: Group_found(S3,a4) is 1; Group_found(S2,a6) and Group_found(S5,a4) is 2; and forth. (Please see the color-coding in Tables 3 and 4).

Third, we compute two subvalues: matching score, and non-matching score. For matching score, if Group_found(s,a) == Group_true(s,a), then we will multiply it with a weight and add it to the score: weights = 1, 0.5, 0.25, 0.125, and 0.1 for the five groups, respectively. This scheme rewards teams that have high accuracy for the top state-action pairs. For the non-matching score, if Group_found(s,a) – Group_true(s,a) == 1 OR Group_true(s,a) – Group_found(s,a) == 1, then we will multiply it with the lower group weight of Group_found(s,a), Group_true(s,a) and add to the score. This is to compensate state-action pairs that miss their true grouping just by one group.

Then we add up the matching and non-matching scores.

| Rank | Leen Dream | Amirite? | Green Kiats | Secret Agent Kiat | Kirspy Nashbrowns | Washington Redskins | Insert Clever Name Here | Kiatten Mittons |
|---|---|---|---|---|---|---|---|---|
| 1 | S3,a4 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 |
| 2 | S2,a6 | S3,a4 | S3,a3 | S3,a3 | S4,a1 | S3,a4 | S3,a4 | S3,a3 |
| 3 | S5,a4 | S3,a3 | S6,a3 | S3,a4 | S6,a1 | S2,a6 | S4,a1 | S3,a4 |
| 4 | S1,a3 | S4,a4 | S3,a4 | S2,a6 | S4,a2 | S2,a2 | S1,a3 | S6,a2 |
| 5 | S5,a2 | S4,a1 | S6,a1 | S6,a3 | S6,a2 | S2,a1 | S1,a4 | S2,a3 |
| 6 | S5,a1 | S5,a1 | S2,a2 | S6,a1 | S4,a3 | S6,a1 | S2,a1 | S1,a2 |
| 7 | S5,a3 | S6,a1 | S5,a2 | S6,a2 | S6,a3 | S2,a5 | S2,a6 | S2,a6 |
| 8 | S3,a2 | S6,a2 | S6,a2 | S6,a4 | S4,a4 | S3,a1 | S2,a2 | S6,a1 |
| 9 | S6,a4 | S3,a6 | S6,a4 | S4,a2 | S6,a4 | S4,a1 | S1,a2 | S1,a3 |
| 10 | S4,a5 | S4,a5 | S6,a5 | S1,a4 | S2,a5 | S5,a1 | S2,a3 | S4,a1 |
| 11 | S3,a6 | S3,a1 | S6,a6 | S4,a1 | S4,a5 | S3,a2 | S2,a4 | S3,a1 |
| 12 | S4,a6 | S4,a3 | S3,a1 | S1,a1 | S6,a5 | S4,a2 | S1,a5 | S3,a2 |
| 13 | S6,a6 | S5,a4 | S3,a2 | S2,a2 | S4,a6 | S5,a2 | S3,a1 | S3,a5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 14 | S1,a1 | S6,a3 | S3,a5 | S1,a2 | S3,a4 | S6,a2 | S5,a1 | S3,a6 |
| 15 | S1,a2 | S4,a2 | S3,a6 | S4,a3 | S3,a3 | S1,a3 | S3,a2 | S2,a4 |
| 16 | S5,a6 | S3,a2 | S1,a3 | S4,a4 | S2,a6 | S2,a3 | S4,a2 | S1,a1 |
| 17 | S3,a1 | S6,a4 | S1,a2 | S4,a5 | S2,a1 | S3,a3 | S5,a2 | S2,a2 |
| 18 | S3,a5 | S5,a5 | S1,a4 | S6,a5 | S1,a1 | S4,a3 | S3,a3 | S2,a1 |

**Table 3.** The ordering of state-action pairs from each team after Round 1. (Only the top 18 state-action pairs are listed) Colors show grouping.

| Rank | Leen Dream | Amirite? | Green Kiats | Secret Agent Kiat | Kirspy Nashbrowns | Washington Redskins | Insert Clever Name Here | Kiatten Mittons |
|---|---|---|---|---|---|---|---|---|
| **1** | S3,a2 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 | S5,a6 |
| **2** | S1,a3 | S3,a3 | S3,a4 | S2,a6 | S4,a1 | S2,a6 | S2,a6 | S2,a6 |
| **3** | S3,a4 | S3,a4 | S3,a3 | S3,a3 | S4,a2 | S6,a2 | S4,a5 | S2,a2 |
| **4** | S5,a6 | S5,a1 | S6,a3 | S6,a2 | S6,a2 | S3,a4 | S3,a4 | S2,a3 |
| **5** | S3,a3 | S6,a1 | S6,a4 | S3,a4 | S4,a3 | S3,a3 | S6,a2 | S3,a3 |
| **6** | S1,a1 | S6,a2 | S2,a6 | S6,a1 | S4,a4 | S6,a3 | S3,a3 | S3,a4 |
| **7** | S3,a1 | S6,a3 | S6,a1 | S6,a3 | S4,a5 | S6,a1 | S6,a3 | S6,a2 |
| **8** | S1,a5 | S6,a4 | S4,a1 | S1,a2 | S6,a4 | S4,a2 | S1,a2 | S1,a2 |
| **9** | S4,a5 | S5,a4 | S5,a2 | S6,a4 | S3,a3 | S1,a4 | S6,a4 | S6,a1 |
| **10** | S2,a6 | S4,a2 | S6,a2 | S1,a4 | S3,a4 | S2,a3 | S6,a1 | S1,a3 |
| **11** | S1,a2 | S4,a5 | S4,a4 | S4,a2 | S2,a6 | S1,a3 | S4,a3 | S4,a1 |
| **12** | S1,a6 | S4,a3 | S4,a5 | S4,a1 | S6,a3 | S2,a2 | S3,a5 | S3,a1 |
| **13** | S4,a6 | S4,a1 | S6,a5 | S1,a1 | S6,a1 | S2,a1 | S3,a6 | S3,a2 |
| **14** | S6,a4 | S5,a5 | S4,a6 | S2,a2 | S2,a1 | S2,a5 | S4,a2 | S3,a5 |
| **15** | S5,a4 | S5,a2 | S6,a6 | S6,a6 | S2,a2 | S3,a1 | S3,a1 | S3,a6 |
| **16** | S5,a1 | S2,a6 | S4,a3 | S6,a5 | S1,a3 | S4,a1 | S3,a2 | S2,a4 |
| **17** | S5,a2 | S2,a2 | S1,a4 | S2,a1 | S1,a4 | S5,a1 | S4,a1 | S1,a1 |
| **18** | S5,a3 | S2,a1 | S3,a1 | S3,a1 | S1,a1 | S3,a2 | S4,a4 | S2,a1 |

**Table 4.** The ordering of state-action pairs from each team after Round 2. (Only the top 18 state-action pairs are listed) Colors show grouping.

Now, we present the more detailed team statistics in Tables 5-7. The number of transactions and rewards were tallied based on the log that our program captured during the Game Day. As shown in Table 5, 4 teams did better than average, and 4 teams performed below average.

| Team Name | #trans | Rewards | Efficiency | Normalized | Order Accuracy | Normalized | Total |
|---|---|---|---|---|---|---|---|
| **Secret Agent Kiat** | 100 | **$66,233.30*** | $662.33 | 1.000 | 2.100 | 0.750 | **1.750*** |
| **Amirite?** | 88 | $37,191.10 | $422.63 | 0.562 | 2.750 | 0.982 | 1.544 |
| **Leen Dream** | 42 | $32,509.20 | $774.03 | 0.491 | **2.800*** | 1.000 | 1.491 |
| **Kiatten Mittons** | 51 | $35,045.70 | $687.17 | 0.529 | 1.450 | 0.518 | 1.047 |
| **Green-Kiats** | 38 | $16,756.60 | $440.96 | 0.253 | 2.175 | 0.777 | 1.030 |
| **Washington Redskins** | 28 | $13,058.60 | $466.38 | 0.197 | 1.675 | 0.598 | 0.795 |
| **Kirspy Nashbrowns** | 25 | $9,806.05 | $392.24 | 0.148 | 1.800 | 0.643 | 0.791 |
| **Insert Clever Name Here** | 22 | $10,622.20 | $482.83 | 0.160 | 1.550 | 0.554 | 0.714 |
| **AVERAGE** | **49.25** | **$27,652.84** | **$541.07** | **0.418** | **2.038** | **0.728** | **1.145** |

**Table 5.** *Statistics of Round 1.* Secret Agent Kiat had the best total score, balancing between rewards and order accuracy, for Round 1. Leen Dream scored the highest order accuracy with 2.8, while Secret Agent Kiat obtained the largest amount of rewards with $66,233.30. * = high value

Table 5 shows only the statistics during Round 2, and *not* the total. There were on average more transactions in Round 2 compared to those in Round 1 (82 vs. 49.25). In terms of Rewards, as expected, Round 2 yielded a higher average than Round 1 ($86,829.14 vs. $27,652.84). This was due to two factors. First, each team's operation, on average, was smoother in Round 2. Second, most team exploited to gain rewards more efficiently ($961.06 per transaction vs. $541.07 per transaction). The average order accuracy for Round 2 was unexpectedly lower than that for Round 1. This was due to teams focusing on exploiting knowledge learned in Round 1 to maximize rewards in Round 2, rather than gaining more accurate knowledge about the Q-table, as described below under "Individual Team Analysis".

| Team Name | #trans | Rewards | Efficiency | Normalized | Order Accuracy | Normalized | Total |
|---|---|---|---|---|---|---|---|
| Kiatten Mittons | 173 | **$222,135.30*** | $1,284.02 | 1.000 | 1.450 | 0.527 | **1.527*** |
| Secret Agent Kiat | 137 | $164,856.70 | $1,203.33 | 0.742 | 2.000 | 0.727 | 1.469 |
| Amirite? | 115 | $104,584.90 | $909.43 | 0.471 | 2.200 | 0.800 | 1.271 |
| Insert Clever Name Here | 45 | $47,614.90 | $1,058.11 | 0.214 | **2.750*** | 1.000 | 1.214 |
| Green-Kiats | 53 | $38,089.40 | $718.67 | 0.171 | 2.400 | 0.873 | 1.044 |
| Kirspy Nashbrowns | 31 | $17,897.15 | $577.33 | 0.081 | 2.450 | 0.891 | 0.971 |
| Washington Redskins | 58 | $58,836.10 | $1,014.42 | 0.265 | 1.350 | 0.491 | 0.756 |
| Leen Dream | 44 | $40,618.70 | $923.15 | 0.183 | 1.200 | 0.436 | 0.619 |
| AVERAGE | 82 | $86,829.14 | $961.06 | 0.391 | 1.975 | 0.718 | 1.109 |

**Table 6(a)** *Statistics of Round 2 (**not** including Round 1's rewards and # transactions).* Kiatten Mittons had the best total score, balancing between rewards and order accuracy, for Round 2. Insert Clever Name Here scored the highest order accuracy with 2.75, while Kiatten Mittons obtained the largest amount of rewards with $222,135.30. * = high value

| Team Name | #trans | Rewards | Normalized | Order Accuracy | Normalized | Total |
|---|---|---|---|---|---|---|
| Kiatten Mittons | 173 | **$223,135.30*** | 1.000 | 1.450 | 0.527 | **1.527*** |
| Secret Agent Kiat | 137 | $166,856.70 | 0.748 | 2.000 | 0.727 | 1.475 |
| Amirite? | 115 | $105,584.90 | 0.473 | 2.200 | 0.800 | 1.273 |
| Insert Clever Name Here | 45 | $47,614.90 | 0.213 | **2.750*** | 1.000 | 1.213 |
| Green-Kiats | 53 | $37,089.40 | 0.166 | 2.400 | 0.873 | 1.039 |
| Kirspy Nashbrowns | 31 | $17,897.15 | 0.080 | 2.450 | 0.891 | 0.971 |
| Washington Redskins | 58 | $55,836.10 | 0.250 | 1.350 | 0.491 | 0.741 |
| Leen Dream | 44 | $40,618.70 | 0.182 | 1.200 | 0.436 | 0.618 |

**Table 6(b)** *Statistics of Round 2 (**not** including Round 1's rewards and # transactions; **including the sales/purchases of Q-tables**).* Kiatten Mittons had the best total score, balancing between rewards and order accuracy, for Round 2. Insert Clever Name Here scored the highest order accuracy with 2.75, while Kiatten Mittons obtained the largest amount of rewards with $223,135.30. * = high value

For Table 6(b), note that Washington Redskins purchased (1) Secret Agent Kiat's Q-table for $2,000, and (2) Kiatten Mittons' Q-table for $1,000; and Green-Kiats purchased Amirite?'s Q-table for $1,000.

Furthermore, though the grand total of the two rounds was not used in our scoring directly, we provide the grand total values for all teams here as a reference in Table 6(c).

| Team Name | #trans | Rewards 1 | #trans | Rewards 2 | #trans Total | Rewards Total |
|---|---|---|---|---|---|---|
| Kiatten Mittons | 51 | $35,045.70 | **173*** | **$223,135.30*** | 224 | **$258,181.00*** |
| Secret Agent Kiat | **100*** | **$66,233.30*** | 137 | $166,856.70 | **237*** | $233,090.00 |
| Amirite? | 88 | $37,191.10 | 115 | $105,584.90 | 203 | $142,776.00 |
| Leen Dream | 42 | $32,509.20 | 44 | $40,618.70 | 86 | $73,127.90 |
| Washington Redskins | 28 | $13,058.60 | 58 | $55,836.10 | 86 | $68,894.70 |
| Insert Clever Name Here | 22 | $10,622.20 | 45 | $47,614.90 | 67 | $58,237.10 |
| Green-Kiats | 38 | $16,756.60 | 53 | $37,089.40 | 91 | $53,846.00 |
| Kirspy Nashbrowns | 25 | $9,806.05 | 31 | $17,897.15 | 56 | $27,703.20 |

**Table 6(c)** *Total rewards and total number of transactions after Round 2.* Kiatten Mittons had the rewards total with $258,181. * = high value

To compute the final score for the Learning Day (50% of the Game Day), we compute the following score for each round:

$$Score = OrderAccuracyNormalized + RewardsNormalized$$

And then we combine both rounds of scores to obtain the final score:

$$FinalScore = 0.4*Score(Round1) + 0.6*Score(Round2)$$

For *OrderAccuracyNormalized*, we normalize each team's order accuracy by the best order accuracy achieved by a team. So, the best team will have its *OrderAccuracyNormalized* = 1.0.

For *RewardsNormalized*, we normalize each team's total rewards (i.e., rewards earned from performing actions + revenue from selling Q-table – cost from purchasing Q-table) with the best rewards earned by a team. So, the best team will have its *RewardsNormalized* = 1.0.

Table 7 shows the result. Overall, Secret Agent Kiat, winners of the first round, scored the highest overall total with 1.585. Amirite?, the second place team and the third place team in Rounds 1 and 2, respectively, narrowly finished second overall They are followed closely by the winner of Round 2, Kiatten Mittons. The Green-Kiats and Insert Clever Name Here finished fourth and fifth, respectively. Finally, Leen Dream, Kirspy Nashbrowns, and the Washington Redskins finished sixth, seventh, and eighth, respectively.

| Team Name | Round 1 Score | Round 2 Score (including sales/purchases) | Final Game Day Score |
|---|---|---|---|
| **Secret Agent Kiat** | **1.750*** | 1.475 | **1.585*** |
| **Amirite?** | 1.544 | 1.273 | 1.381 |
| **Kiatten Mittons** | 1.047 | **1.527*** | 1.335 |
| **Green-Kiats** | 1.030 | 1.039 | 1.035 |
| **Insert Clever Name Here** | 0.714 | 1.213 | 1.014 |
| **Leen Dream** | 1.491 | 0.618 | 0.967 |
| **Kirspy Nashbrowns** | 0.791 | 0.971 | 0.899 |
| **Washington Redskins** | 0.795 | 0.741 | 0.763 |

**Table 7.** Final Game Day scores. Final Game Day Score = 0.4*Round 1 Score + 0.6*Round 2 Score.

## Individual Team Analysis

First, Table 8 shows the learning rate and discount factor used in Round 1 and Round 2 by each team.

| Team Name | Round 1 | | Round 2 | |
|---|---|---|---|---|
| | **Alpha** | **Beta** | **Alpha** | **Beta** |
| **Secret Agent Kiat** | 0.75 | 0.3 | 0.25 | 0.5 |
| **Amirite?** | 1/k (k unique for each state-action pair) | 0.2 | 1/k | 0.2 (0.8 in pregame) |
| **Kiatten Mittons** | Not reported (Pregame: 1 – sqrt(t/6m) where t is the number of minutes that have elapsed since the round began and m is the total number of minutes in the round) | Not reported (Pregame: 0.6) | 0.6 – sqrt(t/3m) (Pregame: 0.5 in the beginning and reduced to 0) | 0.7 (Pregame: Lower beta if standing with relative to class is strong; raise otherwise) |

| | | | | |
|---|---|---|---|---|
| **Green-Kiats** | 0.8 | 0.5 | Not reported (Pregame: decrease by marginal amounts) | Not reported (Pregame: decrease by marginal amounts) |
| **Insert Clever Name Here** | 0.7 (also reported 0.5)<br>(Pregame: 0.5) | 0.5 (also reported 0.3) (Pregame: 0.3) | 0.7 | 0.5 |
| **Leen Dream** | 1/(1.5^t) | 1/n = times visited<br>(Note: Not sure what the above means)<br>(Pregame: 0.3) | 1/(1.5^t)<br>(Midgame: reported 0.5)<br>(Pregame: not discussed) | t = times visited<br>(Note: not sensible)<br>(Midgame: reported ½^t)<br>(Pregame: not discussed) |
| **Kirspy Nashbrowns** | 0.5 | 0.8 | 0.5<br>(Pregame: not discussed) | 0.8<br>(Pregame: not discussed) |
| **Washington Redskins** | 0.8 | 0.5 | 0.4<br>(Pregame: Lower it) | 0.5<br>(Pregame: keep it roughly the same) |

**Table 8.** Learning rates and discount factors used by each team for Round 1 and Round 2.

Before we start looking at teams individually, here is a general sense of the two rounds and the role of the intermission's information sharing.

In general, Round 1 is for exploration, and Round 2 is for a bit more exploitation. That is, Round 1 should be used to explore different state-action pairs. And as a result, one should use a higher learning rate, to emphasize each current transaction and its reward more. The intermission's information sharing should give each team some ideas about how their ordering compares to others. If your team's ordering is very different from others', perhaps your Q-values for these state-action pairs have not converged. If your team's ordering is very similar to others', then perhaps your Q-values have converged. Given that logic, then Round 2 should be more for exploitation if you are confident that your Q-values have converged. In that scenario, using a lower learning rate and a bigger discount factor will help towards that.

But, one critical issue is that what if other teams' orderings are less accurate than yours. Since your confidence in your own Q-values depends on how they match up, what should one do? This is where agent observations and interactions come into play. For example, your team may observe what other teams are doing. Given your observation of other teams' behaviors, you should be able to disregard untrustworthy offers or "description", thereby better utilizing the intermission's "information" sharing to determine your learning rate and discount factor more appropriately.

There are also other factors. Note that for any learning approach to work, in particular for reinforcement learning to work, there must be sufficient learning episodes. In this Game Day, that means each team should secure a lot of transactions in order to better model the stochastic nature of the environment.

Table 9 below shows some correlations among the number of transactions, rewards, and order accuracy values. As expected, the number of transactions and rewards received by each team were highly correlated (0.9175 and 0.9842, respectively, for Rounds 1 and 2). Further, the number of transactions and order accuracy were more correlated in Round 1 when compared to Round 2. And so was the correlation between rewards and accuracy. In general, our intuition is correct. In Round 1, teams used their transactions to explore the various state/action pairs to understand the environment and populate their Q-Tables. Then, exploiting their learned knowledge from Round 1, teams turned to maximizing rewards in Round 2. In particular, in Round 2, teams overall performed many actions exploiting the best state/action sequence at the cost of exploring other paths to improve the Q-Value estimates of suboptimal state/action sequences. In fact, accuracy actually *suffered* in Round 2 (as evidenced by negative correlations between accuracy and both transactions and rewards) since teams performed more actions exploiting what they learned in Round 1. However, this did not hurt the overall rewards earned by teams because most teams correctly identified the optimal state/action pair in Round 1 and found sequences of state/actions leading to this pair's highest reward.

| | **Correlations** | | |
|---|---|---|---|
| | #Trans – Rewards | #Trans – Accuracy | Rewards – Accuracy |
| **After Round 1** | 0.9175 | 0.4635 | 0.3561 |
| **After Round 2** | 0.9842 | -0.2826 | -0.3292 |

**Table 9.** Correlations between number of transactions, rewards, and order accuracy.

Table 10 documents my comments on each team's worksheet and reports. My observations are contextualized on the discussions above. For "Post-Game", I selected some statements from each team's post-game analysis.

| **Team Name** | **Comments** | |
|---|---|---|
| **Secret Agent Kiat** | **Pre-Game** | Had strategies for both rounds: exploration in Round 1 and exploitation in Round 2; mention of utilizing intermission's information sharing; also contingency planning; work distribution; very well prepared |
| | **Round 1 Tracking** | Recorded (very complete) |
| | **Mid-Game** | Report activities, no mid-game observations in terms of changing strategies |
| | **Round 2 Tracking** | Recorded (very complete) |
| | **Post-Game** | **"Our central strategy and centerpiece for both Round One and Two is speed."**<br>**"We found that reducing alpha between Rounds One and Two helped us avoid negating our Round One learning with our exploitation in Round Two."**<br>**"Our preparation, both in the application we built and in our delegation of responsibilities helped us go quickly."**<br>**"… we saw first-hand how a simple algorithm like Q-learning can lead to powerful results."** |
| | **My Observation** | Strategy in Round 1 is maximizing exploration. Application makes R1 recommendations favoring state-action pairs that have not been explored/been explored infrequently. (Not a strict rule – if Q-values stabilizing … begin exploitation) Application makes R2 recommendations favoring reward maximization.<br>They led the game day for the most part but were slowed down by trying to record everything in Round Two. |
| **Amirite?** | **Pre-Game** | Had strategies for both rounds: exploration in Round 1 and exploitation in Round 2; mention of utilizing intermission's information sharing; also contingency planning; work distribution |
| | **Round 1 Tracking** | Recorded (very complete) |
| | **Mid-Game** | Report activities, no mid-game observations |
| | **Round 2 Tracking** | Recorded (very complete) |
| | **Post-Game** | **"We found that in order to fully explore the map, we could not simply rely on** |

| | | |
|---|---|---|
| | | **the algorithm iterating with a low discount rate.”**<br>“We were unable to change the beta parameter without having to reset the Q-value table to random numbers.” |
| | **My Observation** | This team was prepared. Their Round 2 was slowed by their process and also mouse malfunction. |
| **Kiatten Mittons** | **Pre-Game** | Have strategies for both rounds: exploration in Round 1 and exploitation in Round 2; mention of utilizing intermission’s information sharing; also contingency planning; work distribution |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Report activities, mid-game observations |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | **“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.”** |
| | **My Observation** | This team was well prepared. Created an ‘R’ application to automate their calculations and store their Q-table. In Round 2, because of their alpha becoming 0, they were able to put away their app and Q-table and were able to execute actions yielding optimal rewards at a high rate. |
| **Green-Kiats** | **Pre-Game** | Had strategies for both rounds: exploration in Round 1 and exploitation in Round 2; mention of utilizing intermission’s information sharing; also contingency planning; work distribution; not as well prepared, however, as it seems no application/software was written to ensure speed. |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Report activities, no mid-game observations in terms of changing alpha and beta |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | “The rules were very unclear regarding if we need to record specific information and the limitations on records.” “There were a few surprises on game day with rewards to reward values, the ability to purchase matrices” |
| | **My Observation** | The Game Day 1 handout specifically said teams would be allowed to purchase matrices from other teams. Also, the rules were set so that each team should be able to explore different game strategies to adopt during game. For example, the rules indeed were not very clear regarding the need to record specific information. We stated that all information should be recorded. But whether to do it depends on each team’s strategy. |
| **Insert Clever Name Here** | **Pre-Game** | Had a strategy but not explicitly for each round; no mention of utilizing intermission’s information sharing; lack of contingency planning; work distribution not specified; not as well prepared, however, as it seems no application/software was written to ensure speed (though there is an excel spreadsheet) |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Report activities, mid-game observations |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | “Should have reduced learning rate to 0 since we had most learning already done & wouldn’t have had to update values” “Smarter to do learning Round 1 + then exploitation in Round 2, instead of a little of both each round” “**Q-learning is very clear now**” |
| | **My Observation** | This team achieve the highest order accuracy in Round 2 despite much fewer number of actions compared to the others. As a result, it received lower rewards. This team didn’t quite balance exploration and exploitation as well. Not quite sure whether the team had a specific strategy for how to proceed with Round 2 |

| | | |
|---|---|---|
| | | after intermission's deliberation. |
| **Leen Dream** | **Pre-Game** | Had a strategy but not explicitly for each round; no mention of utilizing intermission's information sharing; lack of contingency planning; work distribution not specified; not as well prepared, however, as it seems no application/software was written to ensure speed (though there is an excel spreadsheet) |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Some activities reported; mid-game observations. Made a mistake with their Q-value, resulting in a much higher value. |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | "Confusion breaks down communication when we're in a hurry." "Learning rate becomes way less important when you have good info." "Be careful when typing!" |
| | **My Observation** | This team did not prepare well. The pregame strategies only talked about initial alpha and beta, and contingency planning was not described well and/or without good rationales. The discount rate used was not sensible. Also turned in an incorrectly formatted Q-table (not following instruction). Reporting was not careful. Seemed disorganized. Seemed that this agent did not function very well. |
| **Kirspy Nashbrowns** | **Pre-Game** | Had a strategy but not explicitly/clearly for each round; no mention of utilizing intermission's information sharing; lack of contingency planning; work distribution specified. Not as well prepared. No application/program was written to ensure speed (though there is an excel spreadsheet) |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Some activities reported; no mid-game observations. |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | "If we were to have another learning day, we would try to get a program to make things more efficient". |
| | **My Observation** | This team did not prepare well. The pregame strategies only talked about initial alpha and beta, and contingency planning was not described well and/or without good rationales. Made decisions or observations without sufficient rationales or insights. For example, why did the team decide NOT to change the learning rate before Round 2? |
| **Washington Redskins** | **Pre-Game** | Had strategies for both rounds; mention of utilizing intermission's information sharing; lack of contingency planning; no work distribution specified. Not as well prepared. No application/program was written to ensure speed (though there is an excel spreadsheet) |
| | **Round 1 Tracking** | Recorded (complete) |
| | **Mid-Game** | Bought two Q-tables … "took the consistently high values from their matrices and averaged them into our matrices so we could influence our decisions"; observed that their highest rewards matched own. |
| | **Round 2 Tracking** | Recorded (complete) |
| | **Post-Game** | "Completing the matrix is key to finding he 'Happy paths' as early as possible." |
| | **My Observation** | The team seemed to be plan to purchase others' Q-tables to boost their accuracy. If that's the case, then the team should have started with exploitation as much as possible in Round 1. But this was not so as the team only completed 28 transactions in Round 1. Not as well prepared. |

**Table 10.** My comments and observations of team strategies, worksheets, and reports.

## Lessons Learned

Here are some overall lessons learned.

1. In general, more transactions led to better learning, as shown in the above correlation numbers (Table 9, from Round 1 to Round 2). Thus, acting quickly and efficiently was critical. Teams that were slow in submitting their actions received fewer transactions, leading to poorer performances.
2. Lowering the learning rate or keeping it the same appeared to work better than increasing the learning rate from Round 1 to Round 2 for this MAS environment. In general, increasing the learning rate as time progresses would tend to unlearn what has been learned.
3. Using a high discount factor could have a clamping effect on the learning performance brought on by a high learning rate. This is because looking into the future term essentially incorporates other Q-values into the fray.
4. The information sharing during the intermission was only exploited by a handful of teams. As alluded to earlier, by comparing your ordering with others' could help you decide your learning rate and discount factor. It could also help you decide what actions to choose to perform in Round 2 for a certain state. Teams that had not done well in Round 1 should exploit this to improve in Round 2.
5. Several teams pointed out the nature of a tradeoff at play: ***trying to maximize rewards while trying to maximize the order accuracy***. These two objectives are in a tug-of-war. Maximizing rewards reduces exploration and increases exploitation, and vice versa with maximizing the order accuracy. Several teams had adopted an opportunistic balancing act: if they encountered a "rewarding" good state, they would keep acting on it until it transitioned out, and if they encountered a new state, they would consult the "information shared" (the excel file of all orderings from Round 1) to pick the likely useful action.
6. Teams that were better prepared—that came with the iterative valuation of the Q-learning algorithm and/or a program/application—were ranked higher. As an agent, each team should be observant, adaptive, responsive, and reflective. Not all teams were "responsive" in a timely manner.
7. Note also that the Q-learning or reinforcement learning does *not* tell us which actions to take given a particular state. However, it does *inform* us that up to now, based on our experience, the Q-value of some state-action pairs. This information allows us to carry out our decision making: Should we explore? Should we exploit?

## Game Days League

Here are the League Standings.

| Team Name | Learning Day | Voting Day | Auction Day | Reputation Day (?) | League Standings |
|---|---|---|---|---|---|
| **Secret Agent Kiat** | **1** | | | | **1** |
| **Amirite?** | **2** | | | | **2** |
| **Kiatten Mittons** | **3** | | | | **3** |
| **Green-Kiats** | **4** | | | | **4** |
| **Insert Clever Name Here** | **5** | | | | **5** |
| **Leen Dream** | **6** | | | | **6** |
| **Kirspy Nashbrowns** | **7** | | | | **7** |
| **Washington Redskins** | **8** | | | | **8** |

## Addendum

We ran thousands of iterations given the Tables 1 and 2, with different alpha (learning rate) and beta (discount rate) values, to generate Q-tables. Here we include a table for beta = 0.8 to give you a sense of the Q-value for each state-action pair.

| | | |
|---|---|---|
| S5 | a6 | 10786.2249 |
| S4 | a5 | 8916.7527 |
| S5 | a1 | 8628.9786 |
| S5 | a2 | 8628.9786 |
| S5 | a3 | 8628.9786 |
| S5 | a4 | 8628.9786 |
| S5 | a5 | 8628.9786 |
| S6 | a4 | 8133.4008 |
| S2 | a6 | 7667.849 |
| S3 | a3 | 7247.5822 |
| S3 | a4 | 7216.2336 |
| S6 | a2 | 7134.2779 |
| S4 | a4 | 7133.4008 |
| S4 | a6 | 7133.4008 |
| S1 | a4 | 6798.9062 |
| S6 | a3 | 6798.0645 |
| S6 | a5 | 6506.7194 |
| S6 | a6 | 6506.7194 |
| S6 | a1 | 6439.1237 |
| S4 | a1 | 6149.9427 |
| S2 | a3 | 6134.2779 |
| S2 | a4 | 6134.2779 |
| S2 | a5 | 6134.2779 |
| S4 | a2 | 6125.1762 |
| S1 | a2 | 6067.0352 |
| S4 | a3 | 5953.0268 |
| S2 | a1 | 5897.5384 |
| S2 | a2 | 5834.1807 |
| S1 | a3 | 5831.6858 |
| S3 | a1 | 5798.0645 |
| S3 | a2 | 5798.0645 |
| S3 | a5 | 5798.0645 |
| S3 | a6 | 5798.0645 |
| S1 | a1 | 5786.2249 |
| S1 | a5 | 5439.1237 |
| S1 | a6 | 5439.1237 |