# CA3: Experiment Design and Hypothesis Testing

Joe Anderson, Nov 2021

## 1 Introduction

This report looks to investigate further the comparison of two of the leading public-domain trading algorithms: IBM's GDX and Vytelingum's Adaptive Aggressive traders (AA). My hypothesis looks to extend the results of Cliff and Snashall's 2019 paper: AA Traders Don't Dominate [1], which discovered several circumstances where AA was beaten by GDX. My experiment will do this by introducing a different variety of trading agents into the simulation markets.

While their paper investigated a wide variety of market conditions, only two combinations of agents were tried: AA, ASAD, GDX and ZIP, and AA, ASAD, GDX and ZIC. I expand this by including two more novel agents, Cliff's Giveaway and Shaver algorithms, and assess whether it is one such situation that aids GDX. This gave the final hypothesis:

*GDX is more profitable than AA (with default parameters) in static markets with periodic replenishment, populated by an equal proportion AA, GDX, ZIP, Giveaway, Shaver agents.*

## 2 Experiment and Results

This is tested by running a series of market experiments using BSE, Snashall's implementation of GDX and Ash Booth's implementation of AA. I also used plotting code from the BSE experiment in Week 7. Across these experiments I maintain an equal ratio of each trader, with the buyers mirroring the sellers. I keep also keep the market conditions relatively simple, using static markets inspired by Vytelingum's M1, M2, M3 and M4. M1 is a symmetric market with prices ranging between 10 and 50. M4 is also symmetric, but with a shifted price range to 20 and 60. M2 and M3 are asymmetric, with M2 having a smaller supply price range and M3 vice versa. A supply and demand plot is shown for each market in Figure 1.

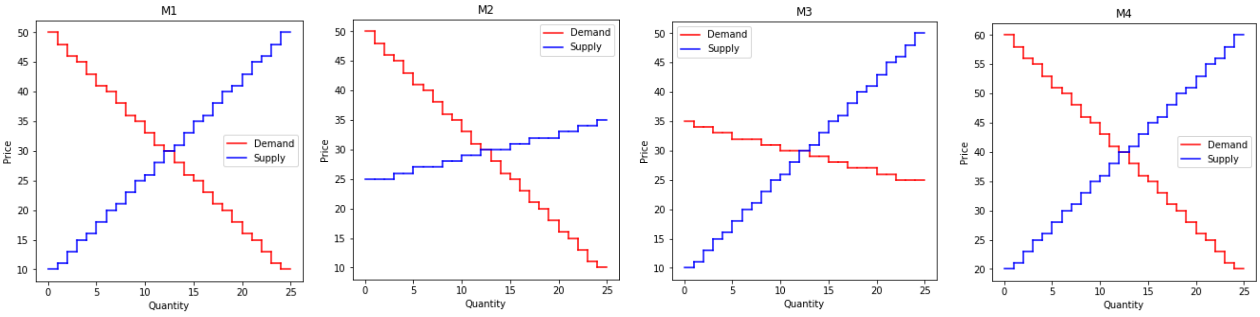

Figure 1: Supply and demand plots for each market

Each of these markets is simulated 200 times, which took a total of 10 minutes of compute time. The results of these simulations are shown in Figure 2. I have omitted the results of the other traders to improve readability (they are not relevant to the hypothesis). However, it is difficult to make many initial conclusions from this graph, except that the algorithms appear evenly matched. To compare the results effectively, statistical tests are run on the data from M1, M2,

[1] Snashall and Cliff, AA Traders Don't Dominate, 2019, https://arxiv.org/ftp/arxiv/papers/1910/1910.09947.pdf

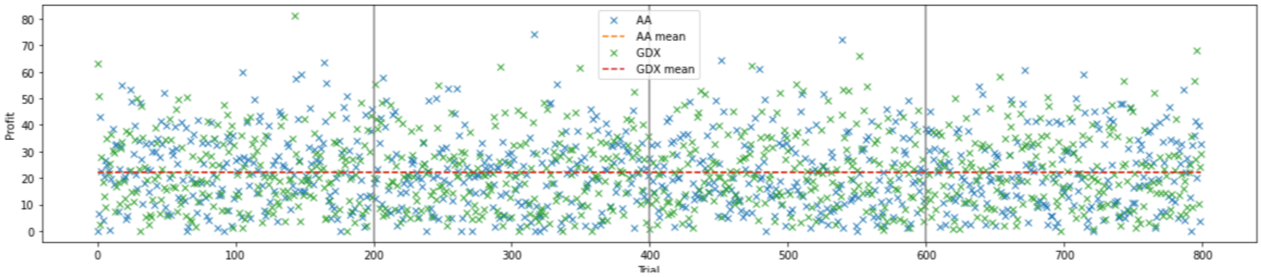

Figure 2: Trader balances for AA and GDX across all experiments

| Market | P-Value |
|---|---|
| M1 | 0.916 |
| M2 | 0.431 |
| M3 | 0.0867 |
| M4 | 0.576 |
| Overall | 0.532 |

Table 1: WMW-U test results.

| Market | P-Value |
|---|---|
| M1 | 0.935 |
| M2 | 0.523 |
| M3 | 0.0956 |
| M4 | 0.541 |
| Overall | 0.597 |

Table 2: Wilcoxon test results.

M3 and M4 individually and as a whole. The first of these tests is the Wilcox-Mann-Whitney U Test, shown in lectures, with significance level 0.05. The hypothesis requires a one-sided test, as we are testing whether the mean score of GDX is greater than that of AA. For this to be rejected, we are looking for p-values less than 0.05. From Table 1 it is clear that there are no significant results in any of the markets, as all the values are greater than the significance level. There was one close-to-significant result in M3, and one very far in the opposite direction (M1).

However, it is important to note that WMW-U assumes independence between the two sets of samples, which is an assumption we may not be able to make in this scenario - it is likely that one agent being more profitable influences the ability of the other agent. For this reason, I also run a Wilcoxon test, which does not assume independence of the two samples. The results from this are shown in Table 2. Again we see no significant results across any of the markets nor overall.

## 3 Conclusion

From these limited experiments, it was not possible to disprove the hypothesis that GDX outperforms AA in these specific market conditions. This is perhaps unsurprising, as in most experiments prior to Snashall, AA had outperformed GDX comfortably. With more time, a more rigorous analysis could be completed that alters more variables and adjusts the parameters of both algorithms.

The null hypothesis may also have been inappropriate – it assumes that our starting position is that GDX is likely the better algorithm. A more open experiment could use a two-sided test that investigates whether the performance of the two algorithms is significantly different. This would avoid accepting one of the two algorithms as better by default.