Project 8 – Strategy Evaluation

Xiaolu Su
xsu73@gatech.edu

Abstract—This report mainly compares the in-sample and out-of-sample performances of a manual strategy and a Q-Learning-based strategy for trading given the same values of three indicators (RSI, MACD, and stochastic oscillator).

1 INTRODUCTION

Despite that the market indicators have become valuable tools for making trading decisions, they rely on a good strategy to generate the best values. Specifically, the project evaluates two strategies developed respectively by human-rules and artificial intelligence. One strategy is a manual strategy using human-defined rules to process the indicator signals, while the other strategy is a Q-Learning reinforcement-based strategy learner that finds the most rewarding policy to utilize the indicator signals. The comparison is enabled by a control group of the same starting money value, commission fee, impact, and the same outputs from three indicators – Bollinger Bands (BB), Relative Strength Index (RSI), and Stochastic Indicator (K). The performances of strategies are also tested on the same in-sample period of the JPMorgan Chase & Co (JPM) trade data from January 1, 2008 to December 31, 2009 and on the same out-of-sample period from January 1, 2010 to December 31, 2011. The only legal trading sizes of an order are -2000, -1000, 0, 1000, and 2000 shares so that the holding can be maintained as -1000, 0, or 1000 shares. Two experiments are conducted to confirm the hypothesis that the Q-Learning strategy learner can secure a higher cumulative return than the manual strategy for both in-sample and out-of-sample periods.

2 INDICATOR OVERVIEW

Four momentum indicators are selected due to their different specialties. Bollinger Bands detects the standard deviation well, Relative Strength Index (RSI) measures the speed and magnitude of the price changes, and Stochastic Indicator inspects a wide range of prices of the day. Altogether they can provide a well-rounded insight into the stock. Their discretizations will be described in more details in Section 4 Table 2.

BB uses a lookback period of 24 days, and its value as an indicator in this project equals the percentage of the difference between price and lower band in the difference of the upper band and lower band.

RSI has a lookback period of the past 14 days, and its outputs are calculated using the adjusted close price. The outputs of RSI are discretized based on the meanings of different ranges of values.

Stochastic Indicator (K) is the percentage of the difference between the current adjusted close price and the lowest trading price within the 14-day lookback period within the difference between the highest trading price within the same 14-day lookback period and the lowest price. Then its signal values (D) equal the 3-day lookback Simple Moving Average of K itself. The equations involved are shown below:

$$K_i = (Adj.Close\_i - Low(14))/(High(14) - Low(14)) * 100$$

$$D = SMA_K(3)$$

The discretization of the Stochastic Indicator is affected by the two required conditions to produce selling or buying signals. One condition is to have $K \leq 20$ or $K \geq 80$. The other condition is to inspect the signs of D-K.

3 MANUAL STRATEGY

3.1 Part 1: In-Sample Manual Strategy vs. Benchmark

For this strategy, the learner takes the pre-discretized values of the four indicators as shown in Table 1. It generates both buying and selling signals by taking the mode of the votes among BB, RSI, and K. While all signals generated by these indicators have been considered, the trades are only created by using the first available buying or selling signal until the next reverse order. While the first order only buys or sells 1000 shares of JPM, the rest of the orders are either in Long (+2000 shares) position or Short (-2000 shares) position. The orders are traded in this way mainly to avoid being misguided by the noises in the stock data and to easily maintain the legal holdings. In addition, the goal of creating this strategy is to exercise being a long-term trader.

To better visualize the performance of Manual Strategy, a Benchmark portfolio is also created, which also started with $100,000 cash but only invested in 1000 shares of JPM and hold this position from the first day of the period. According to Figure 1, Manual Strategy performs generally better than the Benchmark starting from 2008-03 and reached a 21.8% of cumulative return at the end of the period compared to the 1.2% return of the Benchmark based on Table 2. Despite there are some dips around 2008-07, 2008-11, and 2009-09, the strategy is effective overall since it detected the downtrend from 2008-11 to 2009-03.

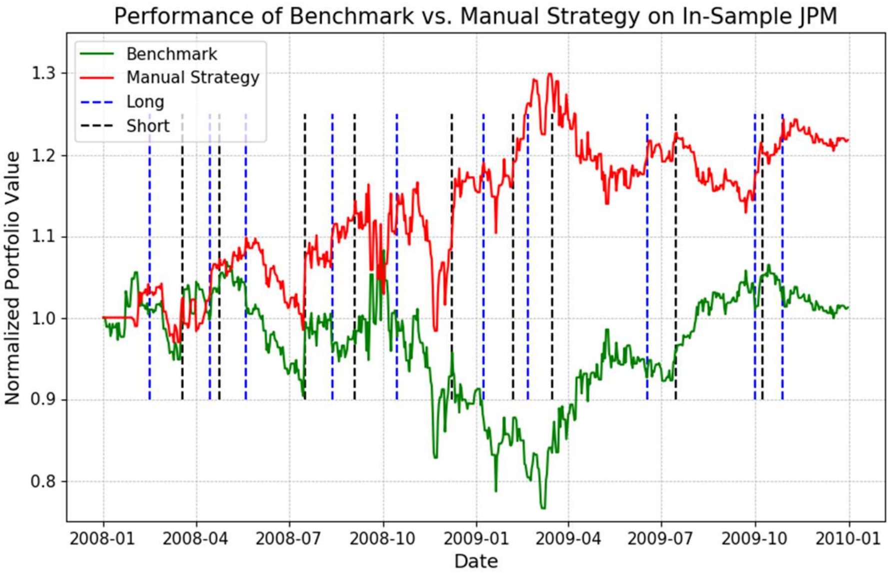

Figure 1—Performance comparison between the Benchmark and Manual Strategy on JPM from 2008, 1, 1 to 2009, 12, 31.

3.2 Part 2: Out-of-Sample Manual Strategy vs. Benchmark

With the same starting money value and rules defined from looking at the in-sample period, Manual Strategy performs not as effective in the out-of-sample period as shown in Figure 2. Even though it still has a higher cumulative return of 5% compared to the Benchmark, the return value is relatively small compared to the in-sample return and the out-of-sample manual portfolio somehow shorted the position when JPM was in an uptrend discerned from the Benchmark portfolio. The effective period of the strategy is most obvious from 2010-02 to 2010-05 and 2010-07 to 2010-12. Possible reasons for the inferior performance can be that the strategy is using the knowledge from the data that is too old for the long-term predictions.

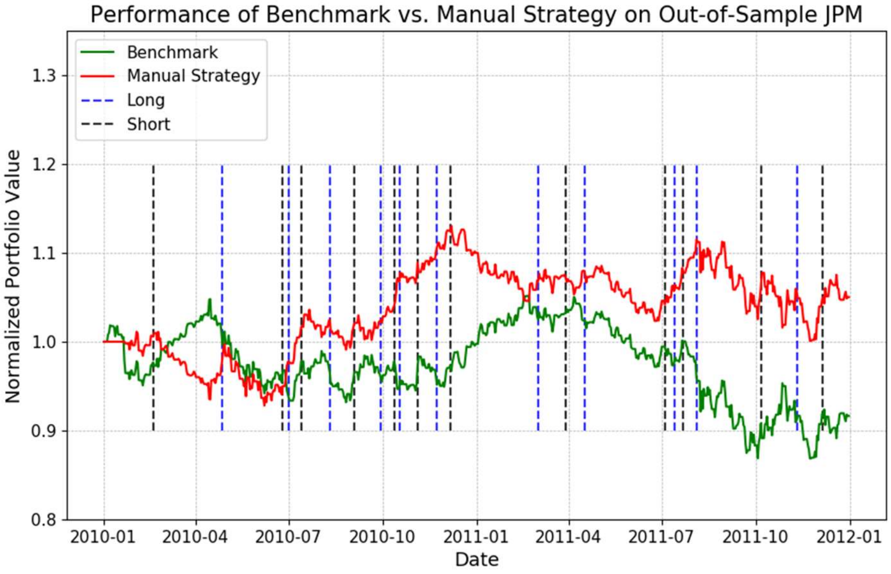

Figure 2 — Performance comparison between the Benchmark and Manual Strategy on JPM from 2010, 1, 1 to 2011, 12, 31.

3.3 Part 3: In-Sample vs. Out-of-Sample

Table 1 — In-sample and out-of-sample metrics of Manual Strategy and Benchmark.

Category	Metrics	Manual Strategy	Benchmark
In-Sample	Cum. Return	0.217850	0.012325
	Stdev (daily)	0.011675	0.014156
	Daily Return	0.000338	0.000117
Out-of-Sample	Cum. Return	0.050313	-0.083579
	Stdev (daily)	0.006280	0.007059
	Daily Return	0.00008	-0.000095

The strategy worked better on the in-sample data partially because it was created with a peek to the future within the in-sample period. The rule itself was designed to generate signals that fit the in-sample period, so when it was used on the unseen, future period, it only achieved a minimal cumulative return with a significance decrease from the in-sample return. Additionally, as shown in Figure 1 & 2, the Benchmark line or JPM itself was more stable in the out-of-sample period than the one in-sample period, which can render the rule learned in the riskier in-sample period less helpful.

As confirmed by the standard deviation (stdev) in Table 2, the effectiveness of the strategy has lowered the risks compared to the Benchmark while maintaining a higher daily or cumulative return. The risks in the in-sample period can be misleading or too unique for a future, flatter period to use the same rules, so the cumulative return can be smaller in the out-of-sample period when there were not as many big slope of risks to be turned into gaining opportunities. The market was likely to have a different agenda going on in 2010 – 2011.

4 STRATEGY LEARNER

The strategy learner is also called a Q-Learner that reinforces good trading decisions while penalizing the bad ones. The learner takes in the discretized values from Bollinger Bands (BB), Relative Strength Index (RSI), and Stochastic Indicator (K) as shown in Table 2. It uses the counts of unique values in each indicator to set up the number of states (n_states). As shown in Table 2, each indicator has 4 different outputs, so n_states = 333 + 1. Then the state value will be based on the current indicator value. For example, when BB=1, RSI=1, K=1, the state=333, which is the last state of all.

Table 2 — Discretizations of all indicators used.

BB	**RSI**	**K**	**Output**	**Description**
Nan	Nan	Nan	-2	Null values
BB < 0	RSI < 35	K < 30 and D-K < 0	-1	Short
0 <= BB <= 1	35 <= RSI <= 60	30 <= K <= 70 or D-K = 0	0	Cash out
BB > 1	RSI > 60	K > 70 and D-K > 0	1	Long

Based on Table 2, the indicators are discretized using different thresholds unique to themselves, and the purposes are also explained in the description. The thresholds are determined based on observations of the tuning results on top of the

JPM stock data in the in-sample period. The values are chosen to make sure the buying signals are generated near a dip of the stock price while the selling signals are generated near a peak of the price. Besides, the indicators thresholds are also tuned to align with the target cumulative return of a lookback period of N=14 days. In this strategy, the legal holding positions are seen as actions, which are Long, Cash out, and Short, so the output values of the discretized indicators also match the position, where Long = 1, Cash out = 0, and Short = -1. The null values are kept to warn the learner from making random decisions without guidance. Since the action in the learner becomes the position, the actual trading orders will be determined as Action = h_prime - h. The overall process can be shown as the pseudo-codes below:

```
While not converged:

      Query initial holding position h_prime by initial state

      Update portfolio (cash, equity, orders) by h=0 & h_prime

      For each day of the stock:

            get the next state s_prime

            get the reward of h_prime

            let h = h_prime and query a new h_prime

            update portfolio with h and h_prime
```

The learner rewards the holding position that leads to positive delayed return (portfolio[t] / portfolio[t – N] – 1) and penalizes the position that leads to a negative daily return. The lookback N=14 is chosen in the tuning process of indicators to keep the information relevant to current position. After each epoch that runs through the entire in-sample period of JPM, a cumulative return was generated to compared with the previous return, and when they are equal, the learning was considered as converged. Once the learner is stabilized, its policy or state querying table would be used directly on the discretized indicator values of the out-of-sample period to generate appropriate signals. The following Experiment 1 will evaluate the performance of the Strategy Learner compared to the Manual Strategy and Benchmark, and it will also develop insights into the influence of impacts in Experiment 2.

5 EXPERIMENT 1

The original hypothesis was that the reinforcement-based learner would achieve a higher cumulative return compared to the Manual Strategy since it has the privilege to utilize reinforcement-based learning that rewards the actions that generates positive returns while penalizing the actions that lead to losses so that it would converge to the optimal policy that knows the best moves in the trading period. Therefore, when comparing the in-sample portfolios, the experiment is assumed to show the line of the Strategy Learner on top, Manual Strategy in the middle, and Benchmark at the bottom.

The indicators involved are BB, RSI, and K as mentioned above and using the same discretization values as shown in Table 2. In the Strategy Learner itself, it finds the combination of $\alpha = 0.3$ $\gamma = 0.5$ $rar = 0.8$ to generate the stable outputs. These parameters represent the learning rate, discount rate, and the random action rate. The experiment set the controlled variables commission = $9.95 and impact = 0.005 for all portfolios. The dyna process was not involved in the QLearner and the random decaying rate was set to 0.9.

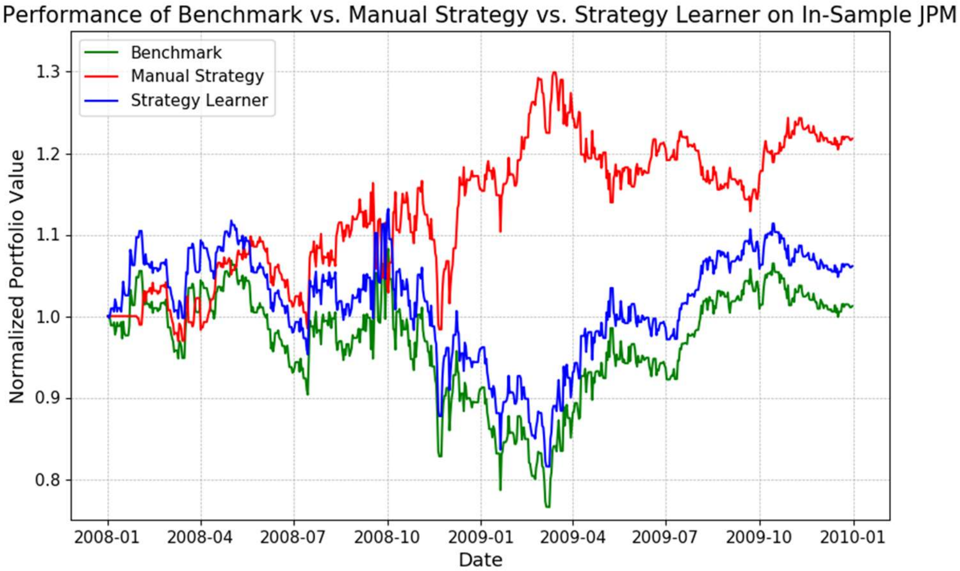

Figure 3—Performance comparisons among the Benchmark, Manual Strategy, and Strategy Learner (Q-Learner) portfolio on JPM from 2008, 1, 1 to 2009, 12, 31.

Unexpectedly, in Figure 3, the strategy learner has converged to the policy that received a lower cumulative return of 7.6% compared to the Manual portfolio. It always goes in the same direction as the trends of JPM. In means that the main position taken during this period has been LONG, and the cumulative return might not be the same for each run with the same QLearner parameters. It could be caused by the initial random actions chosen and led the future states converge to a local optimal. Another reason might be the disagreement among the indicators. Other than the long signals, they have different starting points of short signals and therefore, the learner cannot generate more shorting trades. The feature N is the lookback period to calculate the reward of each action. As it decreased from 14 to 1, while the overall line shape did not change, it was shifted down by around 0.06. There have been multiple tuning trials to adjust the alpha, gamma, and rar for the QLearner in the previous version of Strategy Learner as shown in Table 3 using four indicators. It shows that a lower random action rate can be paired with a medium learning rate and low discount rate to generate the best results with strong stability and relatively higher returns. Future experiments will focus on improving the reward systems and modify the indicator discretization to generate better outputs.

Table 3 — Tuning experiments specific to the combination of Bollinger Bands (BB), Relative Strength Index (RSI), Moving Average Convergence / Divergence (MACD), and Stochastic Indicator (K).

N (Lookback)	α	γ	**rar**	**InSample CR**	**Stable**
4	0.2	0.6	0.8	-0.036	N
2	0.2	0.6	0.8	0.157	N
1	0.2	0.6	0.8	0.351	N
1	0.2	0.1	0.8	0.321	N
1	0.6	0.1	0.6	0.331	N
1	0.5	0.25	0.25	0.372	Y
1	0.5	0.5	0.8	0.370	Y

6 EXPERIMENT 2

As the impact rate discounts the gains or adds interest to the losses, it is somehow a penalty paid to any trade order. The hypothesis of this experiment is that as the impact rate increases, the Strategy Learner is likely to lower its cumulative return and other general performance. For this experiment, it used the same QLearner parameters as Experiment 1 ($\alpha = 0.3$ $\gamma = 0.5$ $rar = 0.8$) and set the commission fee to $0 so that only the impact was changing. It developed three portfolios respectively using impact=0.0005, 0.01, and 0.05 to see the difference. Even though results could differ, as expected, in Figure 4 and Table 4, Strategy Learner with the lowest impact of 0.0005 got the best result. However, as the impact reached 0.01, which is 20 times as large as the lowest impact value, it is not affecting the portfolio performance as much. Their cumulative return difference is only 4.3% and the standard deviation difference is 0.1%. It is only reasonable that either portfolio was not trading frequently and the overall position they held was both SHORT. Unlike the portfolio of impact 0.01, the portfolio with impact of 0.05 was suffering severely from the leverage trades.

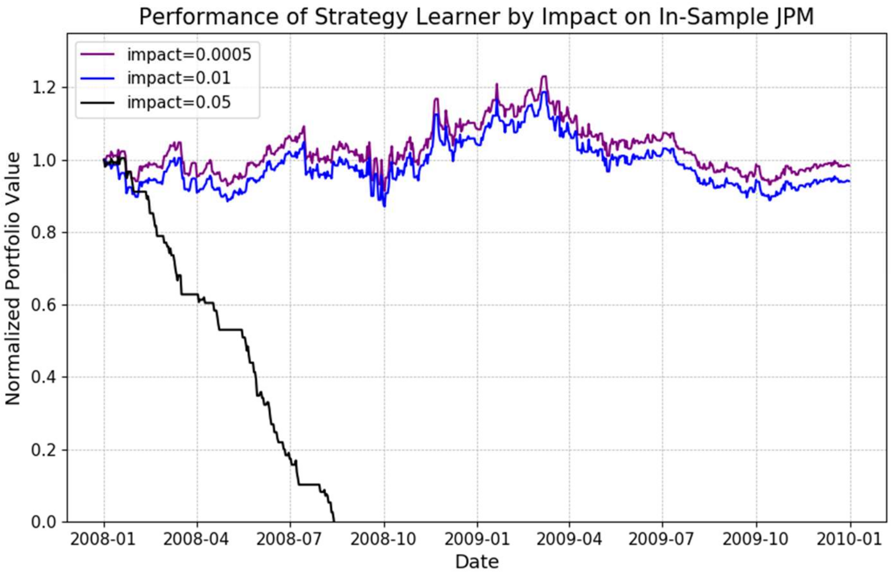

Figure 4—Strategy Learner portfolios from 2008, 1, 1 to 2009, 12, 31 trained with different impact values

The Sharpe Ratios of the three also align with the other two metric, where the best performing portfolio gets the highest Sharpe Ratio, which is the ratio of return to risks. It is still worth noting that the percentage of change to the three metrics was significantly different. While the standard deviation of the impact 0.05 portfolio was only 10 times as large as the impact 0.0005, the difference between their cumulative return could be already up to 320 times of the lowest value. It could indicate that a relatively small change in risks can lead to a big difference in the portfolio performance.

Table 4 — Metrics of in-sample Strategy Learner with different impact values.

Impact	Standard Deviation	Sharpe Ratio	Cumulative Return
0.0005	0.012831	0.074463	-0.016024
0.01	0.013409	0.007711	-0.059065
0.05	0.130869	-0.192205	-3.25502

It can be concluded that, if the fixed penalty like impact is out of a certain range, the QLearner strategy would not be good at handling trades since impact penalizes every action it makes. It could confuse the learner's own reward system and prevent the exploration of more opportunities. And it might be the weakness of the reinforcement-based learning method if more tuning process is not involved. For future experiments, the Strategy Learner will go through multiple trials to find the possibility of balancing out the losses caused by high impact by tuning the parameters like α, γ, rar, lookback period, or it can pursue other combinations of indicators and distinct state numbers.

7 FINAL NOTES

Appended table shows some of the tuning trials. It concludes that for the given period length, it would be advantageous to combine up to 4 indicators or 1200 distinct states despite the inferior cumulative return compared to the manual approach. It would be also preferred to keep a high random action rate and small learning rate to allow more chances for the learner to be trained. While adjusting the lookback period in the calculation of the indicators did not make much difference, it is better keep the lookback period relatively recent before adjusting the thresholds of the indicators. Finally, it is crucial to choose the right sets of actions because a small change of meaning can affect the model seriously.

8 APPENDIX

Table 5 — Tuning experiments. Action "L, N, S" stands for "Long, Do nothing, and Short", while action "L, C, S" stands for "Long, Cash out, and Short".

Indicators	N_states	Lookback	Action	α, γ, r	InSample CR
BB, MACD, K	192	24, 9, 20	L, N, S	0.5, 0.3, 0.9	-0.010
BB, MACD, K	192	24, 9, 20	L, C, S	0.5, 0.3, 0.9	0.011
BB, RSI, K	192	24, 24, 30	L, C, S	0.2, 0.3, 0.9	0.011
BB, RSI, K	336	24, 24, 30	L, C, S	0.2, 0.7, 0.9	0.011
BB, RSI, MACD	192	24, 14, 14	L, C, S	0.2, 0.5, 0.8	0.018
BB, RSI, MACD	336	24, 14, 14	L, C, S	0.2, 0.5, 0.8	-0.021
RSI, MACD, K	192	12, 14, 20	L, C, S	0.2, 0.6, 0.8	0.018
RSI, MACD, K	240	12, 12, 20	L, C, S	0.2, 0.6, 0.8	0.018
BB, MACD, MOM	420	12, 12, 40	L, C, S	0.3, 0.8, 0.8	0.011
BB, MACD, MOM	300	12, 12, 100	L, C, S	0.3, 0.8, 0.8	0.011
BB, RSI, MACD, K	960	30, 14, 12, 20	L, C, S	0.3, 0.8, 0.8	0.022
BB, RSI, MACD, MOM	1200	30, 14, 12, 20	L, C, S	0.3, 0.65, 0.8	0.035
BB, RSI, K, MOM	960	24, 14, 14, 50	L, C, S	0.3, 0.65, 0.8	-0.02
RSI, MACD, MOM, K	1200	14, 9, 20, 14	L, C, S	0.3, 0.65, 0.8	-0.02
BB, MACD, K, MOM	1200	24, 9, 20, 14	L, C, S	0.3, 0.65, 0.8	0.011
ALL	4800	24, 14, 9, 20, 14	L, C, S	0.3, 0.65, 0.8	-0.02