Empirical Evaluation of Multi-armed Bandit Algorithms

ECE8803 Online Decision Making - Course Project Report

Georgia Institute of Technology, Atlanta, Georgia

Supriya Sundar
ssundar47@gatech.edu

Deepa Sree Manogaran
dmanogaran3@gatech.edu

1 Abstract:

While the multi-armed bandit model enjoys popularity in studying the exploration-exploitation trade off in sequential decision-making, there remains ambiguity in the optimal algorithm with respect to regret, confidence and bounds when applying to a specific problem. The multi-armed bandit problem is defined as the determination of choice of arm to pull in a 'K' armed slot machine to maximize profit in a sequence of trials. Many different algorithms exist in theory to this end, but no concrete comparative analysis can be comprehended with respect to a specific dataset or application. This project proposes to identify the optimal algorithm for stock market analysis using empirical evaluation of Upper Confidence Bound(UCB) and Thompson Sampling algorithm.

Upper Confidence Bound(UCB) is favored in financial analysis since it restricts the sampling over time to the actions showing the best performance to maximize the total reward. It progresses from significant exploration to significant exploitation ensuring highest mean reward selected in accordance with the confidence measure. Thompson Sampling is of significance as it gradually refines a model of the probability of the reward for each action and actions are chosen by sampling from this distribution. This gives an estimate for the mean reward value of an action with a measure of confidence for that estimate; allowing optimal action identification faster. This body of work aims to determine the efficacy of these two algorithms in applications of fiscal data.

2 Introduction:

Computing the optimal combination of assets to be held in the portfolio to achieve maximal cumulative return with respect to an agreeable risk measure is the problem defined by portfolio selection. It is a complex problem. Conventional methods initially utilized mean-variance analysis termed as the Modern Portfolio Theory (MPT) which implies choice of allocation that maximizes the expected reward based on the risk associated with variance. Recent times have seen the emergence of sequential portfolio selection models like Cover's universal portfolio strategy [1], Helmbold's multiplicative update portfolio strategy [2]. Current trends observe the power of online decision-making in this regard; application of sequential-decision making with uncertainty to real world problems like portfolio selection, advertising and such. This process involves computation of the optimal trade-off between exploration and exploitation, i.e., finding the balance between naively trying out each machine that may give a possible favorable outcome to iteratively trying out a specific machine when it is known to produce favorable outcome under a given number of trials with given number of machines.

Akin to a single-state Markov decision process (MDP), the stochastic multi-armed bandit (MAB) problem allows provision of a fairly intuitive mathematical framework to study sequential decision-making in the face of uncertainty. Portfolio selection is a dynamic problem on account of the ever changing economic environment. The drastic fluctuations in the market price in this era of multimedia technology makes selection of assets with the goal of maximizing returns while having limited data a complex problem. The multi-armed bandit algorithms facilitates optimization of data and best asset selection in parallel.

An asset is deemed risky if the prices associated with it over a period of time is not relatively constant and this risk needs to be accounted for allocating assets to build the portfolio. The role of the user of the proposed algorithms or the investor is to take into account this risk, analyze the given stock or asset data and identify the optimal one so as to maximize return of investment or reduce the risk long term. Thus, this is a sequential decision-making problem as it requires the investor to constantly monitor and take decisions to increase cumulative reward. The generic regret metric used to analyze the efficacy of these multi-armed bandit algorithms(MAB) is :

$$R_T(X^T) := H_T - \sum_{i=1}^{T} l(x^*; X^T) \tag{1}$$

3 Methodology:

One commonly followed strategy when adding stocks to the portfolio is to choose a subset of best performing stocks from every industry sector in the market. This helps diversify the portfolio while also ensuring the best possible long term rewards. We attempt to model the problem of choosing the best stocks for a portfolio in the multi-armed bandit context. Every stock is an arm in the MAB problem and the algorithms choose one arm/stock every day/time step. Once the arm is sampled, the reward obtained in accumulated and used for the next day's attempts. In a general MAB setting, the rewards of the non-sampled arms are unknown. But since stock market data for a previous day is publicly available, the data can be used for estimating the rewards of the arms that were not sampled. For the purposes of applying the MAB algorithms for stock selection, the algorithms don't make use of the past data of the stocks that were not sampled. The data sets and the definition of a win/loss for a stock is described in the Experiments section.

The two multi-armed bandit algorithms in consideration, Upper Confidence Bound and Thompson Sampling, are described below. The algorithms were coded in Python based on this description. Details of the data sets used and the results from testing the algorithm are described in the consecutive sections.

3.1 Upper Confidence Bound

The Upper Confidence Bound (UCB) algorithm is known for its optimism in the face of uncertainty and was thus presumed to be the optimal choice for the dynamic ever changing stock data.

At any given time, the average rate of reward of a particular machine can be visualized as a point estimate. Each machine is initially presumed to possess a uniform confidence interval and a success distribution. This confidence interval denotes the bounds of success rate distributions, namely the reward value in this case. Based on the confidence interval computed, we can intuitively predict the uncertainty boundary associated with each point estimate as well. This helps us determine the lower and upper boundary of each machine. The interval is the range that is most likely to

consist of the actual success rate distribution of each machine which is unknown to the user at the beginning of the trials.

Since all machines have equal confidence interval at the beginning of the trial, a machine is randomly chosen to be played. Depending on the reward or loss attained, the corresponding machine's confidence interval shifts, shrinking in size or converging from the true success distribution. Thus, whenever an individual machine is tried, its confidence level reduces comparatively and this ensures that each machine is tried at least once without any bias. The algorithm is designed to choose the upper bound in every trial. Based on the "Upper Confidence Bound" of each machine, the highest value machine is selected to be tried in the subsequent trial. These steps are tried iteratively till all the machines have been tried and sufficient observations have been made to determine the maximal confidence bound with certainty.

The general flow of the algorithm iteration is depicted below:

Algorithm 1: UCB

Input: $X, K = |X|$

$t \leftarrow 1$

$\forall x \in X : \hat{\mu_x} \leftarrow 0$

while $t \leq T$ **do**

  $x_t \leftarrow argmax_{x \in X} \left(\hat{\mu}_x + \beta \sqrt{\frac{\ln t}{n_x}} \right)$

  play arm x_t and acquire u_t

  update $\hat{\mu}_x$

Figure 1: Upper Confidence Bound Algorithm

3.2 Thompson Sampling

Upper Confidence Bound (UCB) algorithm tackles the multi-armed bandit problem by choosing the next action based on the current averages of the rewards from the previous actions. Thompson Sampling algorithm approaches the multi-armed bandit problem in a Bayesian perspective by building a probability model for each arm based on the obtained rewards and then samples from the posterior probabilities to choose the next action.

In addition to the key difference above, the UCB algorithm always learns from scratch. In most of the practical applications of multi-armed bandit algorithms, there is usually some known information that could be used to sample the arms and converge to the best arm faster. This could prove useful when choosing the best stocks to add to your portfolio as faster the selection, higher the total rewards obtained by adding the stock to the portfolio. Thompson sampling algorithm uses the priors available for every arm and then computed the posterior probabilities based on which the arms are sampled at random.

The algorithm is briefly described in figure 2. At the start of the experiment, all the arms are assigned corresponding priors, which can be same or different for every arm. For every time step, the posterior probability is computed based on Beta priors. The arms are randomly sampled from the resulting posterior distribution and the arm with the maximum posterior probability is pulled. Once pulled, the priors for the arm is updated as an increment to the parameters of beta distribution. The rewards are accumulated over each pull and is used to measure the overall performance of the algorithm for each experiment.

Algorithm 1 Thompson sampling algorithm

Input: Prior $\boldsymbol{Q}^{(a)}$ on arm a for $a = 1, \ldots, K$
for $t = 1, \ldots, T$ **do**
  Compute posterior distribution $\boldsymbol{Q}_t^{(a)}$ on μ_a from observed samples
  Sample $(\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{K,t})$ from the posterior distributions $\boldsymbol{Q}_t^{(a)}$
  Pull arm $A_t = \arg\max_{a \in \{1,\ldots,K\}} \mu_{a,t}$, and observe reward G_{t,A_t}.
end for

Figure 2: Thompson Sapling Algorithm

4 Experiments, Observations and Discussions:

The algorithms described above were first tested with a randomly generated data set - a stream of 0's and 1's - to verify the implementation. Once the observations matched the expectations, we moved on to testing the algorithms on the stock data. The historic data for stocks were obtained from Google Finance website. For the purposes of this experiment, five stocks of Fortune500 companies were randomly chosen for a period of five years. The algorithms choose a new stock once a day. This time step, however, doesn't impact how the algorithms perform and can be tuned when necessary.

The expected observation is that the MAB algorithms pick the stock with more "wins" over the last five years. If a stock increases by at least by one dollar, it is declared as a win. This definition of a "win" can be altered according to the strategy of the user. A better strategy can accelerate the learning and lead to a faster stock selection. For instance, if a "win" is defined as any amount gained in a day, most stocks will have a similar number of wins/losses and can lead to multiple trials of the sub-optimal arms before converging to the best stock.

4.1 Observations on a random data sequence

4.1.1 UCB

Figure 3 showcases the summative reward of five arms considered over a period of fifty runs.Random mean values were initially taken. The linear rise in the curve with increase in the number of steps proves that the algorithm has explored and is exploiting the optimal arm to maximize rewards.

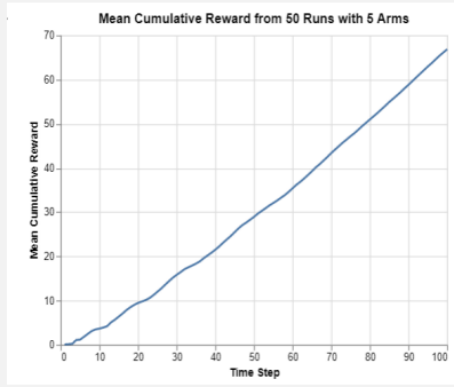

Figure 3: UCB applied to random data

4.1.2 Thompson Sampling

Figures 4 and 5 show the gradual convergence of the Thompson sampling algorithm on the most efficient arm. The random sequence was generated as a Bernoulli sequence with user defined probability for every arm. In Figure 4, the initial priors were set to 1 for all five arms and each of the arms result in a success with probabilities 0.1, 0.5, 0.3, 0.7, 0.25 respectively. In Figure 5, the initial priors for arms 1 and 5 were set to 0.833 and 0.67. In both cases, the algorithm converges to the most efficient arm within few trials as possible.

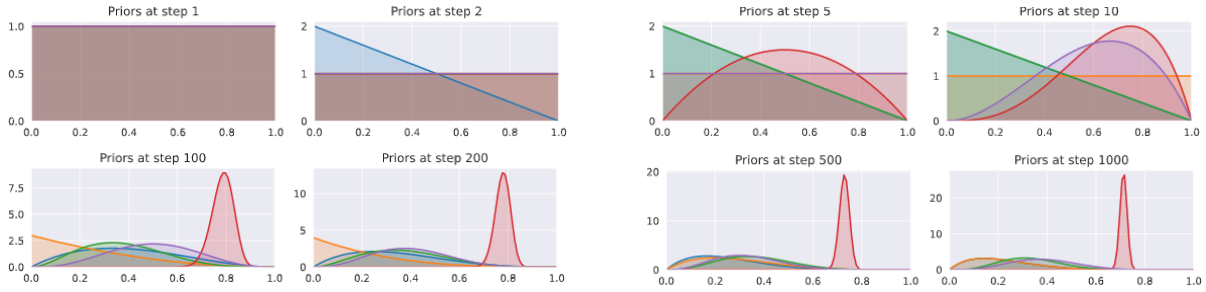

Figure 4: Priors from Thompson Sampling on random generated data with equal priors

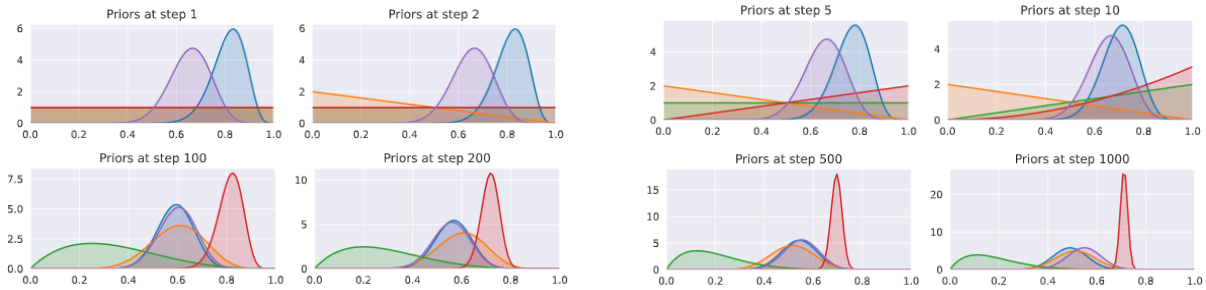

Figure 5: Priors from Thompson Sampling on random generated data with different priors

4.2 Observations on stock data

4.2.1 UCB

Figure 6 portrays the cumulative and individual rewards achieved over the said time period. The upper bound is computed as per the formula illustrated in the methodology section and this bound is applied to the shuffled data. It is observed that Amazon stock has the highest reward value amidst the five company stocks.

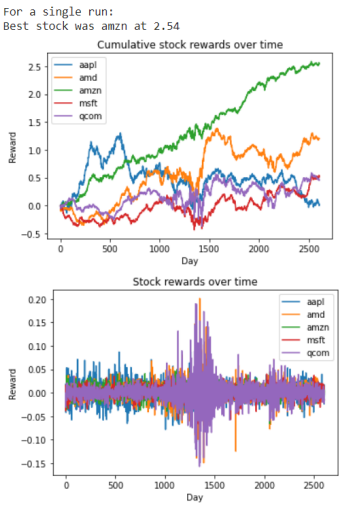

Figure 6: UCB applied to stock data

4.2.2 Thompson Sampling

Figure 7 summarizes the trend in the chosen stocks over the period of five years. Using the definition of rewards and "wins" defined in this section, the priors were plotted against time steps for the processed stock data. "Green" stock with most one dollar increments per day over the time period is chosen as the most optimal arm by the Thompson Sampling algorithm.

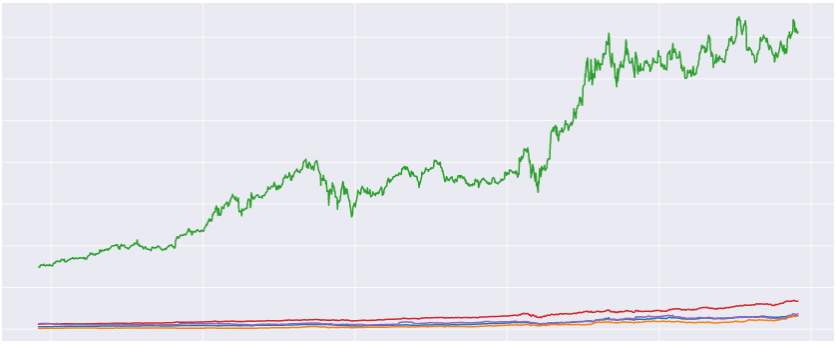

Figure 7: Stock trend over past five years

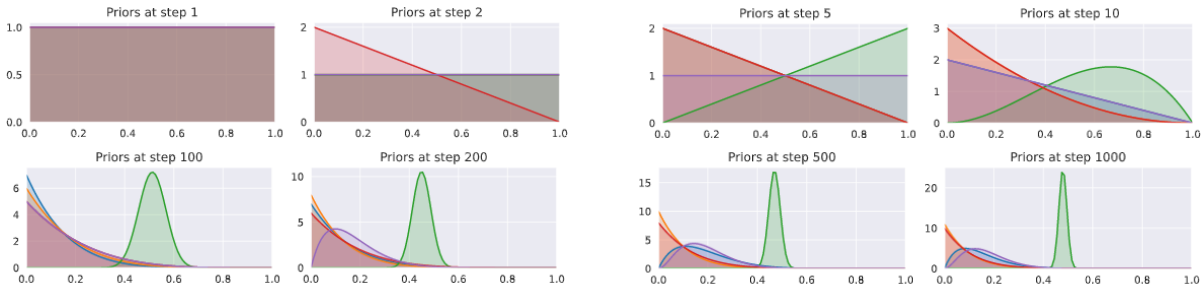

Figure 8: Priors from Thompson Sampling on stock data with same priors

5 Conclusion:

The multi-armed bandit problem has been scrutinized as a numerical model for sequential decision-making under uncertainty via this project. The application of this problem in the finance sector has been analyzed in depth using Thompson Sampling and Upper Confidence Bound algorithms. The opening and closing, low and high value of each individual stock was analyzed over a period of five years. This individual pricing of the stock was modeled as a curved graph to analyze the growth or fall in rate and the general trend of data has been observed. It was inferred that the software company stock, Amazon has performed consistently better than the other four semiconductor companies stocks, namely Apple, AMD, Microsoft and Qualcomm when considered over the period from 2017 till 2022. Post application of Thompson Sampling to the priors and posteriors generated, the posteriors sampled and the arm picked at random recursively led us to the optimal stock data. Parallelly, Upper Confidence Bound was applied to the data and the company with the maximal reward return was computed from the upper confidence bound.

Thus, the optimal exploration-exploitation trade off achieved can be discerned from the optimal choice of company stock in the end. The balance between reducing risk and maximizing returns is also observed. It is concluded that these results agree well with the finance industry data provided there are no drastic external factors affecting it. As the stock market is extremely volatile in nature, risk awareness, knowledge of past data, and optimal prediction methodology plays an important role in portfolio compilation.

6 Future Work:

Analysis of trends in the stock market prior to and after the Covid pandemic of 2020 could be analyzed in depth to get a more practical understanding of consumer interests in stock investments. This would provide us with a more novel portfolio that can be deployed practically in current times in the industry.

We may scrutinize an unstable market environment where stock values are affected by numerous external factors instead of following a straight forward stochastic process. Random Matrix Theory and Transfer Entropy has been utilized by Junior and Mart [3] to illustrate the effect of news articles on stock data.

Portfolio selection can also be done by finding the optimal combination of algorithms in place of a single one. The possibility of using a particular sampling approach and a greedy approach beyond a certain point of time or threshold may be explored in depth.

7 References:

[1] T.M. Cover. Universal portfolios. Math. Finance, 1(1):1-29, 1991.

[2] Helmbold, D. P., Schapire, R. E., Singer, Y., and Warmuth, M. K. 1996. On-line portfolio selection using multiplicative updates. In Proceedings of the International Conference on Machine Learning. 243–251.

[3] Junior LS, Mart AM.2017. Correlations and flow of information between The New York Times and Stock Markets. (http://arxiv.org/abs/1707.05778)

[4] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285-294, 1933.

[5] Tor Lattimore and Csaba Szepesvari. Bandit algorithms. Cambridge University Press, 2020.