MAFS5370-Project-1 RL-in-asset-allocation

Ding Zihang

Github Link: https://github.com/DarvinDing/MAFS-Project-1

1. Problem

Consider the discrete-time asset allocation example in section 8.4 of Rao and Jelvis. Suppose the single-time-step return of the risky asset as Y_t=a, $prob$=p, $and\ b$, $prob$=(1−p). Suppose that T =10, use the TD method to find the Q function, and hence the optimal strategy.

2. Analytical Solution

The problem setting is we have Wt at discrete time steps t = 0,1, …T-1. Assume the single-time-step discount factor is γ and that the Utility of Wealth at the final time step t = T is given by the following CARA function:

$$U(W_T) = \frac{1 - \exp(c * W_T)}{c}\ for\ fixed\ c\ != 0$$

Our goal is to maximize the expected value:

$$E\left[\gamma^{\mathrm{T-t}}\left(\frac{1 - \exp(-a * W_T)}{a}\right)|(t, W_t)\right]$$

where xt ∈ R, γT−t and c are constants. This is equivalent to maximizing:

$$E\left[-\frac{\exp(-a * W_T)}{a}\middle|(t, W_t)\right]$$

Therefore, we formulate this problem as a MDP with the following components:

- Target Wt is the Wealth at time t
- State st ∈ St at timestep t = 0,1, …, T − 1
- Random Variable Yt determines each step of the risky return, where Yt has the binomial distribution:

$$Y\text{t}= \{a, probability\ of\ p$$
$$\{\text{b}, probability\ of\ 1\text{-}p$$

- Constant riskless return r
- Action at ∈ At at timestep t = 0,1, …, T − 1 determines the risky investment xt. Hence, the investment in the riskless asset at time t will be Wt – xt
- A deterministic policy at time t (for all t =0, 1, ..., T −1) is denoted as πt, optimal deterministic policy at time t (for all t = 0,1,...,T − 1) is denoted as π∗ t
- MDP reward is $E[-\frac{exp(-a*W_T)}{a}\ |\ (t, W_t\ )\ ]$

Then the wealth Wt+1 at t = t+1 has the below form:

Wt+1 = (Wt − xt) · (1 + r) + xt · (1 + Yt) = xt · (1 − r) = xt · (Yt − r) + Wt · (1 + r)

We denote the Value Function at time t (for all t = 0, 1, . . ., T − 1) for a given policy π = (π0, π1, . . . , πT−1) as: $V_t^{\pi}(\mathrm{W_t}) = \mathrm{E}_{\pi}\left[-\frac{\exp(-a\mathrm{W^T})}{a}\middle|(\mathrm{t},\mathrm{W_t})\right]$

We denote the Optimal Value Function at time t (for all t = 0, 1, . . ., T − 1) as:

$$V_t*(Wt) = \max_{\pi} V_t^{\pi}(Wt) = \max_{\pi}\{E_{\pi}[-\frac{\exp(-aW_T)}{a}|(t,Wt)]\}$$

The Bellman Optimality Equation is:

$$V_t*(Wt) = \max_{xt} Q_t^*(Wt,x_t) = \max_{xt}\{E_{Y_{T-1}}[-\frac{\exp(-aW_T)}{a}|(t,Wt)]\}$$

For all t = 0, 1, …, T-2 and

$$V_{T-1}*(W_{T-1}) = \max_{x_{T-1}} Q_{T-1}^*(W_{T-1},x_{T-1}) = \max_{xT-1}\{E_{Y_{T-1}}[-\frac{\exp(-aW_T)}{a}|(t,Wt)]\}$$

Where $Q*$ is the Optimal Action-Value function.

We make a guess for the functional form of the Optimal Value Function as:

$$V_t*(Wt) = -b_t \exp(-ct\cdot Wt)$$

where bt, ct are independent of the wealth Wt for all t = 0, 1, . . ., T − 1.

Then we can express the Bellman Optimality Equation as:

$$V_t*(Wt) = \max_{xt}\{E_{Y_t}[-b_{t+1}\exp(-c_{t+1}\cdot W_{t+1})]\}$$

Which can be further expressed using

Wt+1 = (Wt − xt) · (1 + r) + xt · (1 + Yt) = xt · (1 − r) = xt · (Yt − r) + Wt · (1 + r)

As

$$V_t*(Wt) = \max_{xt}\{-b_{t+1}(p*\exp\left(-c_{t+1}\cdot\left(xt\cdot(a-r)+Wt\cdot(1+r)\right)\right)+(1-p)\exp(-c_{t+1}\cdot(xt\cdot(b-r)+Wt\cdot(1+r)))\}$$

We then infer the functional form for $Q^*(Wt,xt)$ in terms of b_{t+1} and c_{t+1}

$$Q_t*(Wt,x_t) = -b_{t+1}(p*\exp\left(-c_{t+1}\cdot\left(xt\cdot(a-r)+Wt\cdot(1+r)\right)\right)+(1-p)\exp(-c_{t+1}\cdot(xt\cdot(b-r)+Wt\cdot(1+r)))$$

To find the maximum of $Q_t*(Wt,x_t)$, we set the partial derivative to 0:

$$\frac{\partial Q_t*(W_t,x_t)}{\partial x_t} = 0$$

After derivation, we get $x_t^* = \frac{1}{c_{t+1}(a-b)} In(\frac{(a-r)p}{(r-b)(1-p)})$

We substitute this maximized value into the Bellman Optimality Equation :

$$V_t * (Wt) = -b_{t+1}(p * \exp\left(-\frac{a-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right) + (1 -$$

$$p) \exp\left(-\frac{b-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right) * \exp(-c_{t+1}(1+r)W_t)$$

We get the equation for bt and ct:

$$\text{bt} = \text{bt+1} \cdot p * \exp\left(-\frac{a-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right) + (1-p) \exp\left(-\frac{b-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right)$$

$$\text{ct} = \text{ct+1} \cdot (1 + \text{r})$$

This gives the new optimal policy:

$$x_t^* = \frac{1}{c_t(a-b)(1+r)^{T-t-1}} In(\frac{(a-r)p}{(r-b)(1-p)})$$

And the new optimal value function:

$$\text{Q}_\text{t}^*(W_t, x_t) = -p * [\exp\left(-\frac{a-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right)$$

$$+ (1-p) \exp\left(-\frac{b-r}{a-b} In\left(\frac{(a-r)p}{(r-b)(1-p)}\right)\right)]\text{^}(T-t)$$

$$* \exp(-c(1+r)^{T-t} W_t)$$

3. Reinforcement Learning

In this section, we will apply the two simple TD method, Q learning and SARSA to provide solution to the optimal policy problem.

3.1 Q-Learning

Environment Setup

The environment for the portfolio optimization algorithm is structured to model wealth dynamics over a specified number of time steps. Key components include:

- **Risk-Free Rate**: A fixed return rate for safe investments.
- **Risky Asset**: Returns modeled using a binomial distribution, introducing upward and downward movements in asset value.
- **Discretized Wealth Levels**: The wealth is divided into 50 discrete levels, where N is the maximum portfolio value. This discretization facilitates the management of a large state space.

- **Allocation Actions**: There are 11 discrete actions representing the proportion of wealth allocated to the risky asset, ranging from 0 to 1.

Q-Value Update Formula

The Q-learning update formula used in this algorithm is:

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma max\ Q(S',A') - Q(S,A)]$$

Where:

- A is the learning rate.
- R is the received reward after taking action A in state S.
- γ is the discount factor, influencing the present value of future rewards.

Key Design Points

Exploration Decay

The algorithm implements an exploration rate decay technique over time. This is crucial in reinforcement learning for balancing exploration (trying new actions) and exploitation (using known rewarding actions). The exploration rate ϵ starts high (e.g., 1) and gradually decays to a lower limit (e.g., 0.001 or 0.05) over a specified number of episodes. This approach allows the agent to explore the action space initially while focusing on exploiting learned strategies as training progresses.

Function Approximation

To efficiently manage the large state space, the algorithm employs function approximation through state and action neighborhoods. Updates to Q-values are spread to neighboring state-action pairs within a defined range (e.g., ±5 wealth indices, ±1 action index). This generalization technique allows the algorithm to share information across similar states and actions, thereby enhancing learning efficiency and accelerating convergence.

Metrics Tracking

During the training process, the algorithm tracks multiple performance metrics, including:

- **Theoretical Discrepancy**: Measures the difference between the expected and actual outcomes.
- **Learning Signals**: Captures the fluctuations in learning progress, such as temporal difference errors.
- **Terminal Utilities**: Records the utility at the end of episodes, reflecting the effectiveness of the strategy.

- **Risk Exposures**: Assesses the level of risk taken in allocations, aiding in the evaluation of the strategy's risk-return profile.

Result Analysis:

1. Convergence Test

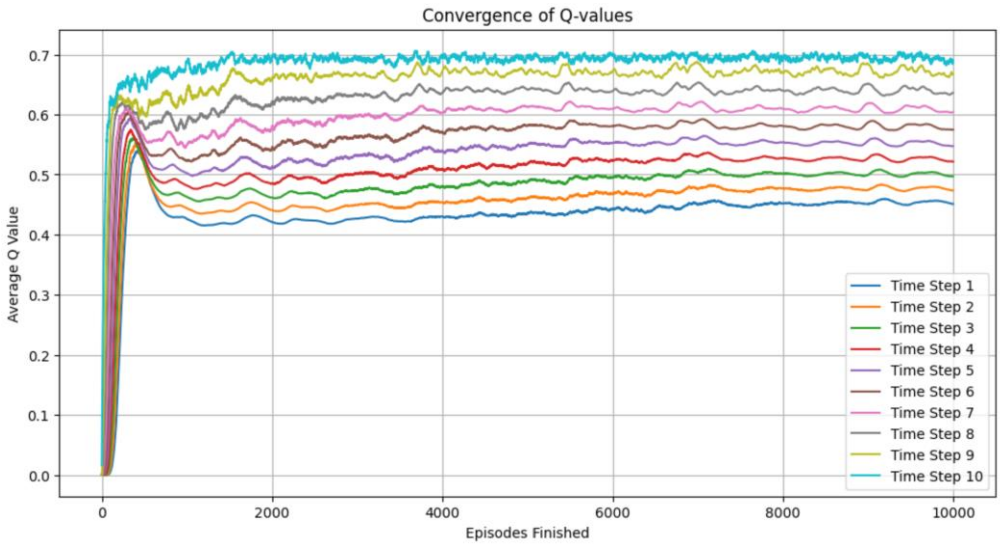
This section analyzes how the Q-values converge during the training of a reinforcement learning agent tasked with optimizing a portfolio.

We used the *train_investment_strategy_with_convergence_test* function to train the reinforcement learner. This involved setting up the portfolio simulator and the learner with specific parameters, such as the risk-free rate, risk aversion coefficient, and the distribution of a risky asset.

Key Parameters:

- *Investment Horizon: 10 time steps*
- *Risk-Free Rate: 2%*
- *Risk Aversion Coefficient (γ): 0.3*
- *Total Training Episodes: 10,000*
- *Learning Rate (α): 0.1*
- *Discount Factor (γ): 0.95*
- *Exploration Rate (ε): Started at 1.0, later reduced to a minimum of 0.01*

Results



The plot shows that the average Q-values stabilize as the training progresses, indicating that the reinforcement learner effectively adapted its strategy. The Q-values for all time steps trend toward stability, suggesting that the learner has successfully adjusted to the investment environment.In the early episodes, we see significant fluctuations in Q-values, which gradually decrease. This reflects the learning process as the agent finds optimal allocations.The initial high exploration rate allows the agent to try various

actions, leading to the fluctuations seen in early episodes. As exploration decreases, the agent focuses on what it has learned, resulting in the observed convergence.

2. Optimal Test

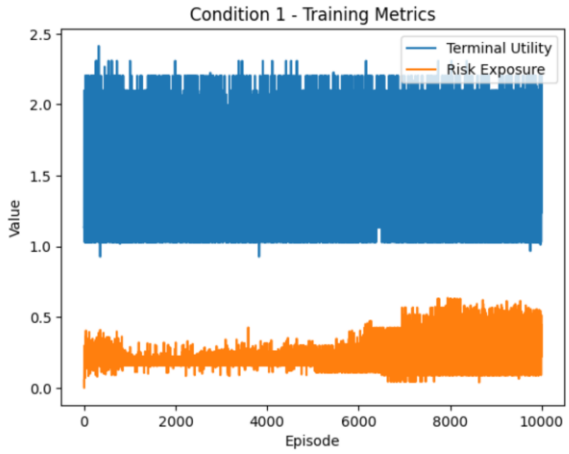
In this section, we investigate the performance of the reinforcement learning agent by simulating wealth changes under various market conditions. Three distinct scenarios were defined, each characterized by different risk-return profiles for the risky asset. The aim is to observe how the learned policy influences wealth evolution over time.

We evaluated the agent's performance under the three conditions, For each condition, we set up a *PortfolioSimulator* and a *ReinforcementLearner* using the specified parameters. After training the investment strategy, we analyzed the learned policy and simulated wealth changes using the learned policy over multiple iterations.

1. Condition 1:
    - Parameters: a=0.05, b=−0.05, p=0.7, r=0.02
    - Expected Behavior: Risk-free asset preferred.

Result:

```
Wealth Change Simulations:
Simulation 1 - Wealth Change: [1.2121212, 1.3131313, 1.3131313, 1.4141414, 1.5151515, 1.6161616, 1.5151515, 1.5151515, 1.6161616, 1.6161616, 1.6161616]
Simulation 2 - Wealth Change: [1.7171717, 1.8181819, 1.8181819, 1.8181819, 1.919192, 1.919192, 1.919192, 1.919192, 2.020202, 2.020202, 2.121212]
Simulation 3 - Wealth Change: [1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101]
Simulation 4 - Wealth Change: [1.5151515, 1.5151515, 1.5151515, 1.6161616, 1.6161616, 1.7171717, 1.8181819, 1.919192, 1.8181819, 1.8181819, 1.8181819]
Simulation 5 - Wealth Change: [1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101, 1.010101]
```

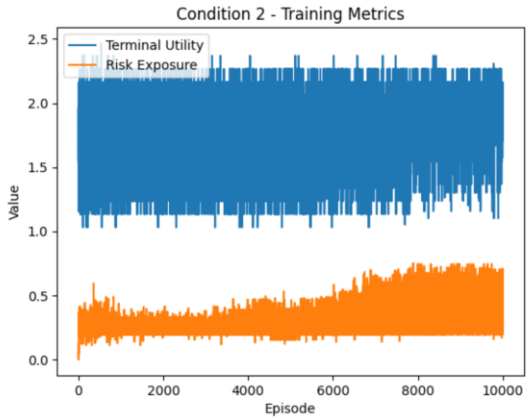


In Condition 1, the results indicated a strong preference for the risk-free asset, reflected in a stabilized average terminal utility of approximately 1.64 throughout the training episodes. The low risk exposure, shown by minimal fluctuations in the graph, reinforced the agent's cautious strategy, allocating only about 0.06 of its resources to the risky asset. Wealth change simulations demonstrated slight increases, maintaining stability around the initial value, which aligns with the risk-averse approach expected under these parameters.

2. Condition 2:
    - Parameters: a=0.07, b=−0.05, p=0.7, r=0.02
    - Expected Behavior: Risky asset preferred due to higher expected return.

Result:

```
Wealth Change Simulations:
Simulation 1 - Wealth Change: [1.010101, 1.1111112, 1.2121212, 1.3131313, 1.3131313, 1.4141414, 1.5151515, 1.6161616, 1.6161616, 1.7171717, 1.7171717]
Simulation 2 - Wealth Change: [1.1111112, 1.2121212, 1.3131313, 1.4141414, 1.5151515, 1.6161616, 1.6161616, 1.7171717, 1.8181819, 1.8181819, 1.8181819]
Simulation 3 - Wealth Change: [1.1111112, 1.2121212, 1.1111112, 1.2121212, 1.3131313, 1.4141414, 1.5151515, 1.6161616, 1.7171717, 1.8181819, 1.8181819]
Simulation 4 - Wealth Change: [1.7171717, 1.7171717, 1.8181819, 1.919192, 1.919192, 1.8181819, 1.919192, 1.919192, 1.919192, 1.919192, 1.919192]
Simulation 5 - Wealth Change: [1.6161616, 1.5151515, 1.6161616, 1.6161616, 1.6161616, 1.7171717, 1.8181819, 1.8181819, 1.919192, 1.919192, 1.919192]
```

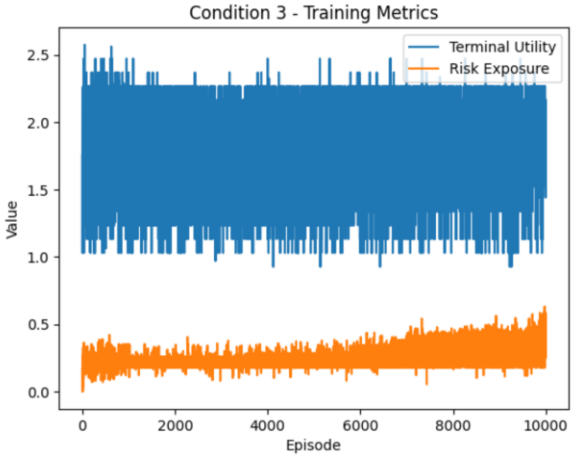


In Condition 2, the findings demonstrated the agent's shift toward the risky asset, as expected due to the higher anticipated returns. The average terminal utility stabilized around approximately 1.80, indicating improved performance compared to Condition 1. The risk exposure showed more significant fluctuations, reflecting the agent's increased willingness to allocate resources to the risky asset, which is evident in the orange line of the graph. Wealth change simulations revealed more pronounced increases, with final wealth levels reaching up to around 1.99, highlighting the benefits of the agent's risk-taking strategy.

3. Condition 3:
    - Parameters: a=0.07, b=−0.05, p=0.9, r=0.02
    - Expected Behavior: Risky asset preferred due to higher probability of profit.

Result:

```
Wealth Change Simulations:
Simulation 1 - Wealth Change: [1.2121212, 1.3131313, 1.4141414, 1.5151515, 1.6161616, 1.7171717, 1.7171717, 1.8181819, 1.8181819, 1.919192, 2.020202]
Simulation 2 - Wealth Change: [1.6161616, 1.7171717, 1.8181819, 1.8181819, 1.8181819, 1.919192, 1.919192, 1.919192, 1.8181819, 1.919192, 2.020202]
Simulation 3 - Wealth Change: [1.4141414, 1.5151515, 1.6161616, 1.7171717, 1.7171717, 1.8181819, 1.8181819, 1.919192, 2.020202, 2.121212, 2.121212]
Simulation 4 - Wealth Change: [1.7171717, 1.8181819, 1.919192, 1.919192, 1.919192, 2.020202, 2.020202, 2.020202, 2.121212, 2.121212, 2.121212]
Simulation 5 - Wealth Change: [1.7171717, 1.8181819, 1.919192, 1.919192, 1.919192, 2.020202, 2.020202, 2.020202, 2.121212, 2.121212, 2.121212]
```



In **Condition 3**, the results showcased the agent's strong preference for the risky asset, driven by the higher probability of profit. The average terminal utility stabilized around

approximately 1.92, indicating a notable improvement from the previous conditions. The risk exposure metric exhibited greater variability, reflecting the agent's aggressive allocation toward the risky asset, as seen in the fluctuations of the orange line in the graph. Wealth change simulations revealed substantial increases, with final wealth levels reaching up to around 2.12, underscoring the effectiveness of the agent's risk-seeking strategy.