Reinforcement Learning with Financial Applications

Assignment #1

GU Zhihao

MAFS: 5370 Homework Assignment

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

March 10, 2025

Problem 1

We are given wealth W_0 at time 0. At each of discrete time steps labeled $t = 0, 1, \ldots, T-1$, we are allowed to allocate the wealth W_t at time t to a portfolio of a risky asset and a riskless asset in an unconstrained manner with no transaction costs. The risky asset yields a random return Y_t satisfying

$$Y_t = \begin{cases} m, & \text{with prob. } p, \\ n, & \text{with prob. } 1-p, \end{cases}$$

over each single time step. The riskless asset yields a constant return denoted by r over each single time step (for a given $r \in \mathbb{R}$). We assume that there is no consumption of wealth at any time $t < T$, and that we liquidate and consume the wealth W_T at time T. So our goal is simply to maximize the Expected Utility of Wealth at the final time step $t = T$ by dynamically allocating $x_t \in \mathbb{R}$ in the risky asset and the remaining $W_t - x_t$ in the riskless asset for each $t = 0, 1, \ldots, T-1$. 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:

$$U(W_T) = \frac{1 - e^{-aW_T}}{a}, \quad \text{for some fixed } a \neq 0.$$

Suppose that $T = 10$, use the Temporal-Difference method to find the Q function, and hence the optimal strategy.

Solution. The problem is to maximize, for each $t = 0, 1, \ldots, T-1$, over choices of $x_t \in \mathbb{R}$, the value

$$\mathbb{E}\left[\gamma^{T-t} \cdot \frac{1 - e^{-aW_T}}{a} \Big| (t, W_t)\right].$$

Since γ and a are constants, this is equivalent to maximizing for each $t = 0, 1, \ldots, T-1$, over choices of $x_t \in \mathbb{R}$, the value

$$\mathbb{E}\left[\frac{-e^{-aW_T}}{a} \Big| (t, W_t)\right]$$

Since we denote the random variable for the single-time-step return of the risky asset from time t to time $t+1$ as Y_t which satisfies

$$Y_t = \begin{cases} m, & \text{with prob. } p, \\ n, & \text{with prob. } 1-p, \end{cases}$$

for all $t = 0, 1, \ldots, T-1$, we have

$$W_{t+1} = x_t \cdot (1 + Y_t) + (W_t - x_t) \cdot (1 + r) = x_t \cdot (Y_t - r) + W_t \cdot (1 + r).$$

The MDP Reward is 0 for all $t = 0, 1, \ldots, T-1$ and for $t = T$, it should be

$$\frac{-e^{-aW_T}}{a}.$$

Thus, by applying Temporal-Difference method to $Q_t^\pi(W_t)$ and $Q_{T-1}^\pi(W_{T-1})$, we should have the iteration:

$$Q_t^*(W_t) \leftarrow \underbrace{Q_t^*(W_t) + \alpha \cdot \left(Q_{t+1}^*(W_{t+1}) - Q_t^*(W_t)\right)}_{(1-\alpha)Q_t^*(W_t)+\alpha Q_{t+1}^*(W_{t+1})}, \quad t = 0, 1, \ldots, T-2.$$

$$Q_{T-1}^*(W_{T-1}) \leftarrow \underbrace{Q_{T-1}^*(W_{T-1}) + \alpha \cdot \left(\frac{-e^{-aW_T}}{a} - Q_{T-1}^*(W_{T-1})\right)}_{(1-\alpha)Q_{T-1}^*(W_{T-1})-\frac{\alpha e^{-aW_T}}{a}}, \quad t = T-1.$$

Finally, we set all the parameters as: $\gamma = 0.95$, $a = 0.5$, $\alpha = 0.3$, Y_t has a positive return $m = 0.15$ with probability $p = 0.6$, negative return $n = -0.05$ with probability $1 - p = 0.4$, $r = 0.03$, $W_0 \in [-200, 400]$. Detailed algorithm is written in the code.