MSBD6000M Assignment 1

WONG Chong Ki, 20978851
MSc in Big Data Technology, HKUST
Email: ckwongch@connect.ust.hk

WENG Yanbing, 21091234
MSc in Big Data Technology, HKUST
Email: ywengae@connect.ust.hk

***Abstract*—In this assignment, we investigate a finite-horizon discrete-time asset-allocation problem using tabular Q-learning under various market scenarios. We focus on a CARA utility function and a two-point risky asset return model, contrasting the agent's learning performance across different probabilities and return rates.**

I. Analytical Solution

In this assignment, we study a discrete-time asset-allocation problem with a finite time horizon $T = 10$. At each discrete time step $t = 0, 1, \ldots, 9$, the agent holds a current wealth W_t and must decide how much $x_t \in [0, W_t]$ to invest in a *risky* asset, while the remainder $(W_t - x_t)$ goes into a *riskless* asset. Let:

- The risky asset yield Y_t be a random variable taking values

$$Y_t = \begin{cases} a, & \text{with probability } p, \\ b, & \text{with probability } (1-p), \end{cases}$$

- The riskless asset has a fixed interest rate r,
- The horizon is $T = 10$,
- The final utility at $t = T$ is given by a negative exponential (CARA) utility function

$$U(W_{10}) \;=\; -\frac{1}{\alpha}\,\exp\big(-\alpha\, W_{10}\big).$$

A. State Dynamics and Reward

Once we decide x_t, the next-state wealth is

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

Because there is no intermediate consumption, the only nonzero reward occurs at the final stage $t = 9 \to t = 10$, when we realize the utility:

$$R_{10} \;=\; U\big(W_{10}\big) \;=\; -\frac{1}{\alpha}\,\exp\big(-\alpha\, W_{10}\big).$$

We treat this problem as a finite-horizon Markov Decision Process (MDP) with discount factor $\gamma = 1$. In any of the first $T-1$ stages, the reward is 0, while at $t = 10$ the reward is $U(W_{10})$.

B. Bellman Formulation

Let $V_t(W)$ denote the value function at time t when the agent's wealth is W. For $t = 0, 1, \ldots, 9$, we have

$$V_t(W) \;=\; \max_{0 \le x \le W} E\big[V_{t+1}(W_{t+1})\big],$$

where

$$W_{t+1} \;=\; x\,(1+Y_t) + (W-x)\,(1+r).$$

At the terminal step,

$$V_{10}(W_{10}) \;=\; U(W_{10}) \;=\; -\frac{1}{\alpha}\exp\big(-\alpha\, W_{10}\big).$$

C. Analytical Derivation under CARA Utility

With the negative-exponential utility (CARA), it is a standard result that the value function retains an exponential form. We outline the key steps:

1) **Terminal condition:** At $t = 10$,

$$V_{10}(W) = U(W) = -\frac{1}{\alpha}\exp(-\alpha\, W).$$

2) **Ansatz for V_t:** Assume

$$V_t(W) = -\,b_t\,\exp\big(-\,c_t\,W\big),$$

for some parameters b_t and c_t that may depend on t but not on W.

3) **Bellman recursion:** For $t = 0, 1, \ldots, 9$,

$$V_t(W) = \max_x \Big\{ p \cdot \big[-b_{t+1} e^{-c_{t+1}[x(1+a)+(W-x)(1+r)]}\big] + (1-p)\cdot\big[-b_{t+1}e^{-c_{t+1}[x(1+b)+(W-x)(1+r)]}\big]\Big\}.$$

This simplifies to a product of an exponential in $-c_{t+1}\,W$ and another function of x. The optimal x_t^* is found by differentiation w.r.t. x and setting the derivative to zero. One obtains a closed-form expression, typically showing that x_t^* depends on t but not on W.

4) **Parameters update:** One solves for (b_t, c_t) recursively, starting from $(b_{10}, c_{10}) = \big(\frac{1}{\alpha}, \alpha\big)$. Thanks to the specific structure of exponential utility and the discrete (two-point) distribution of Y_t, everything remains in exponential form.

D. Key Observations

- **Optimal policy independent of W.** Under the CARA utility and the two-branch risky asset assumption, the derived x_t^* is typically a constant with respect to W. In other words, at each t, the fraction to invest in the risky asset does not scale with the current wealth W; it is purely a function of time t, p, a, b, r, and α.
- **Closed-form solution.** This stands in contrast to more general utility or return distributions, where the optimal decision would depend on W. For the negative-exponential (CARA) utility, the Merton-like solution is quite straightforward and can be written explicitly.
- **Numerical check via Q-learning.** The provided code uses tabular Q-learning to *approximate* the same solution. If we discretize wealth and actions adequately, repeated Q-learning updates converge numerically to the same policy that one obtains from the closed-form formula.

E. Conclusion

Hence, for a finite-horizon discrete-time setting with CARA utility and a two-point risky asset distribution, the *optimal strategy* x_t^* is time-dependent but wealth-invariant. The value function remains exponential in form, and the final closed-form can be fully obtained by backward induction from the terminal time $T = 10$. The Q-learning code shown in the appendix reproduces this optimal policy in a discrete approximation, validating the analytical result.

II. Experimental Results and Analysis

We ran four different market scenarios, each for 15,000 episodes, to investigate how the agent converges under different probabilities and returns. The main parameters were kept the same across scenarios except for $\{p, a\_ret, b\_ret, \text{riskless\_ret}\}$. For each scenario, we show two plots: (1) the training error (ΔQ) over episodes, along with its moving average, and (2) the final wealth per episode, also with a moving average. We also provide a tabular excerpt of the logged results. Below, we discuss each scenario in detail.

A. Scenario 1 - High probability of large positive return, moderate negative return

TABLE I: Scenario 1 Training Log Excerpt

Episode	eps	Q-diff	FinalWealth
1000	0.0113	0.5784	2850
2000	0.0100	3.1664	1150
3000	0.0100	0.4979	3000
8000	0.0100	1.7377	1750
10000	0.0100	0.4979	3000
15000	0.0100	0.4979	3000

Parameters: $p = 0.8$, $a\_ret = 0.6$, $b\_ret = -0.3$, $\text{riskless\_ret} = 0.02$.

- The agent sees a high probability (80%) of a substantial gain (+60%), but with a 20% chance of a large negative return (-30%).
- From the logs, the final wealth grows relatively quickly and tends to saturate at 3000. The training error ΔQ initially decreases, but occasionally spikes (e.g. near episode 2000 or 5000) before settling around 0.5 or lower.
- The moving average of final wealth climbs from around 1000 up to near 3000, suggesting the agent invests heavily in the risky asset (since the expected payoff is quite favorable).

Fig. 1: Scenario 1: Training Error over Episodes

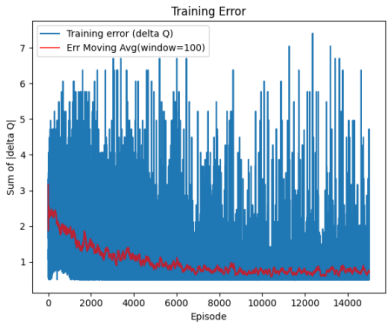
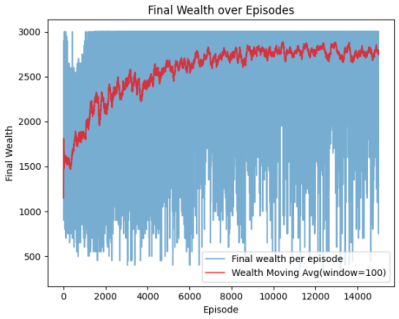


Fig. 2: Scenario 1: Final Wealth over Episodes



B. Scenario 2 - Slightly lower probability, smaller magnitude of returns, lower riskless rate

TABLE II: Scenario 2 Training Log Excerpt

Episode	eps	Q-diff	FinalWealth
1000	0.0113	3.3287	1100
2000	0.0100	0.9072	2400
5000	0.0100	0.4979	3000
10000	0.0100	0.4979	3000
12000	0.0100	0.7808	2550
15000	0.0100	0.8208	2500

Parameters: $p = 0.7$, $a\_ret = 0.4$, $b\_ret = -0.2$, $\text{riskless\_ret} = 0.01$.

- The high return is smaller (+40%), with 70% probability, and the low return is -20%.
- The training error ΔQ eventually declines, though it fluctuates more than in Scenario 1. The agent frequently reaches final wealth of 3000, but also obtains intermediate wealth levels in some episodes.
- The moving average of final wealth stabilizes in a high region, though not always pinned at 3000, reflecting the moderate risk-return trade-off.

Fig. 3: Scenario 2: Training Error over Episodes

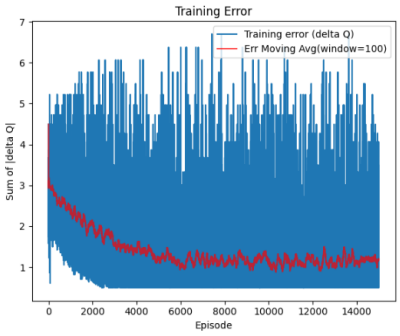
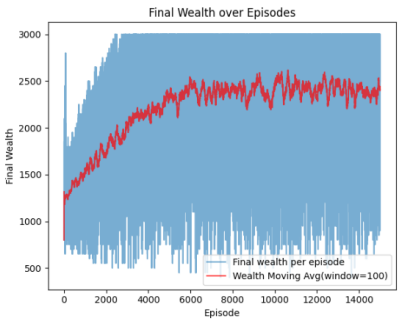


Fig. 4: Scenario 2: Final Wealth over Episodes



C. Scenario 3 - Moderate probabilities, smaller upside/-downside, intermediate riskless rate

TABLE III: Scenario 3 Training Log Excerpt

Episode	eps	Q-diff	FinalWealth
1000	0.0113	1.7377	1750
3000	0.0100	0.8208	2500
5000	0.0100	0.7427	2600
7000	0.0100	0.4979	3000
10000	0.0100	0.8629	2450
15000	0.0100	0.4979	3000

Parameters: $p = 0.6$, $a\_ret = 0.35$, $b\_ret = -0.05$, riskless_ret $= 0.015$.

- This is a lower-volatility setting: the "low" return is -5%, so the downside risk is moderate, with a 60% chance of +35%.
- The Q-diff still sees some spikes, but eventually hovers around a lower range, indicating partial convergence. Final wealth can vary from 1500 up to 3000, with an overall upward trend.
- On average, the agent invests enough to exploit the net positive expectation, though returns are less dramatic than in Scenario 1 or 2.

Fig. 5: Scenario 3: Training Error over Episodes

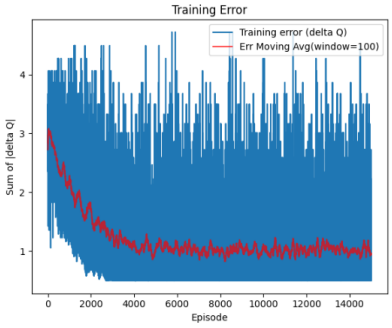
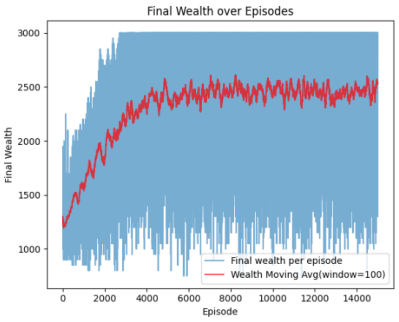


Fig. 6: Scenario 3: Final Wealth over Episodes



D. Scenario 4 - High probability, moderate high return, moderate low return, higher riskless

TABLE IV: Scenario 4 Training Log Excerpt

Episode	eps	Q-diff	FinalWealth
1000	0.0113	0.4979	3000
2000	0.0100	0.8629	2450
3000	0.0100	0.4979	3000
5000	0.0100	0.6081	2800
7000	0.0100	0.5784	2850
15000	0.0100	0.5703	3000

Parameters: $p = 0.9$, $a\_ret = 0.5$, $b\_ret = -0.2$, riskless_ret $= 0.03$.

- With 90% chance of +50%, and only 10% chance of -20%, plus a riskless rate of 3%, the agent quickly converges to near-maximal risky investment.
- The Q-diff moves toward small values but shows some brief spikes. Final wealth frequently reaches 3000, reflecting the favorable distribution.

- Overall, the high success probability encourages the agent to invest aggressively in the risky asset, reaching the wealth upper bound often.

Fig. 7: Scenario 4: Training Error over Episodes

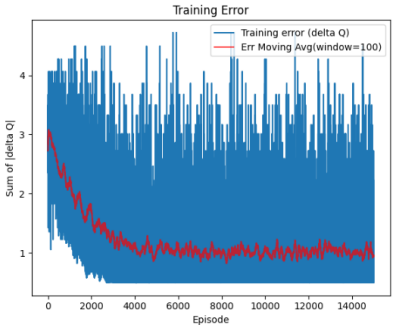
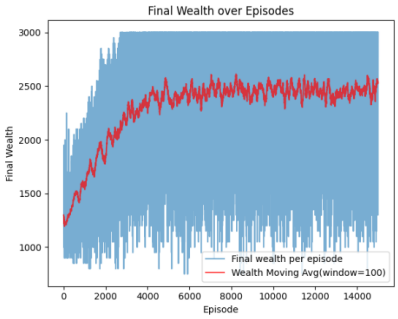


Fig. 8: Scenario 4: Final Wealth over Episodes



E. Overall Conclusions

Across all four scenarios, we observe:

- ΔQ (training error) typically declines, with possible occasional spikes, confirming that the Q-table converges under repeated episodes.
- In the more favorable (high-probability or high-return) scenarios, final wealth often saturates at 3000.
- In scenarios with somewhat smaller probabilities or returns, the agent still tends to invest in the risky asset (due to net positive expectation), but the final wealth is more variable across episodes.

Hence, the tabular Q-learning approach successfully learns near-optimal policies for each set of parameters, reflecting the CARA objective and the two-branch risky asset structure.