Assignment 1 Report:

1. Discrete Case:

We use three methods: Q-Learning, Sarsa(λ), and Sarsa(0) to find the Q function and optimal policy.

Q-Learning:

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

Sarsa(λ):

$$\delta \leftarrow R + \gamma Q(S',A') - Q(S,A)$$
$$Q(S,A) \leftarrow Q(S,A) + \alpha\delta E(S,A)$$
$$E(S,A) \leftarrow \gamma\lambda E(S,A)$$

Sarsa(0):

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

For Q-Learning, each update is the Q value that takes action in the current state, while TD(λ) updates the current state's value function. Q-Learning does not use Eligibility Traces.

TD(λ) uses Eligibility Traces to track the frequency of access to each state, thus weighing the contributions of current and previous states to value functions.

Then Sarsa lets agent learn Q(s,a) instead of V(s), and becomes an on-policy TD control method: Sarsa(λ).

Sarsa(0) just considers one-step, and the eligibility trace for all states except the current state is 0.

For the ***State*** space, which is the wealth at time t, we assume it is discrete including all discrete outcomes(wealth) after each time step. For the ***Action*** space, which indicates the type of holding asset at one time step including two values: 0 and 1 (0: hold risk-free asset, 1: hold risky asset).

2. Q function approximation with semi-gradient Sarsa

We use Episodic Semi-gradient Sarsa for Estimating action-value function introduced in Section 10.1 from RL Book. We try on action-value function approximation via linear model, polynomial model with order 2, exponential model. Besides, after we try on analytical solution, we derive the 'true' action-value function by assuming the form for optimal value function and inferring the optimal action value function from Bellman Optimality Equation.

For the ***State*** space, which is the wealth at time t, we assume it is continuous in [0,1]. For the ***Action*** space, which is the quantity of investment in the risky asset, we discretize it into *ACTION_SPLIT* portion from [0,1].

Linear Model

$$\boldsymbol{Q(w_t, x_t) = q_0 \cdot w_t + q_1 \cdot x_t + q_2}$$

Polynomial Model with order 2

$$\boldsymbol{Q(w_t, x_t) = q_0 \cdot w_t + q_1 \cdot w_t^2 + q_2 \cdot x_t + q_3 \cdot x_t^2 + q_4}$$

Exponential Model

$$\boldsymbol{Q(w_t, x_t) = e^{q_0 \cdot w_t + q_1 \cdot x_t + q_2}}$$

Analytical Solution

$$\boldsymbol{Q_t(w_t, x_t) = -b_{t+1} e^{c_{t+1} \cdot [x_t(r-b) - w_t(1+r)]} \cdot \left[1 - p + p \cdot e^{c_{t+1} \cdot x_t(a-b)}\right]}$$

3. Analytical Solution

We derive the analytical solution given the risky asset follows the known distribution. Here are some significant results, and the whole procedure could be found in attached file. Assume the value function be:

$$V_t(w_t) = -b_t \cdot e^{-c_t \cdot w_t}$$

$$Q_t(w_t, x_t) = -b_{t+1} e^{c_{t+1} \cdot [x_t(r-b) - w_t(1+r)]} \cdot \left[1 - p + p \cdot e^{c_{t+1} \cdot x_t(a-b)}\right]$$

$$b_T = \frac{1}{a}, c_T = a$$

$$X_t^* = \frac{1}{-c_{t+1} \cdot (\alpha - \beta)} In\left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p}\right)$$

Where **r** denotes the risk-free return, **a** be the risk aversion coefficient.

Plug X_t^* into value function we derive:

$$b_{t+1} \cdot \left[p \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p}\right)^{\frac{\alpha-r}{\alpha-\beta}} + (1-p) \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p}\right)^{\frac{\beta-r}{\alpha-\beta}}\right] = b_t \ for\ t = T-2,..1$$

$$c_{t+1} \cdot (1+r) = c_t \ for\ t = T-2,..1$$

Considering the T-1 period, we derive:

$$b_{T-1} = \left[p \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p}\right)^{\frac{\alpha-r}{\alpha-\beta}} + (1-p) \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p}\right)^{\frac{\beta-r}{\alpha-\beta}}\right] \cdot \frac{1}{a}$$

$$c_{T-1} = \mathrm{a} \cdot (1+r)$$

As a result:

$$b_t = \left[p \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p} \right)^{\frac{\alpha-r}{\alpha-\beta}} + (1-p) \cdot \left(\frac{r-\beta}{r-\alpha} \cdot \frac{p-1}{p} \right)^{\frac{\beta-r}{\alpha-\beta}} \right]^{T-t} \cdot \frac{1}{a} \quad for\ t = T-1,..1$$

$$c_t = a \cdot (1+r)^{T-t} \quad for\ t = T-1,..1$$

4. Further models based on Analytical Solution

Test on Simple Case, T=1:

Since we have known the analytical solution of value function, we could conduct analysis with shorter period , like T = 1, to check that whether Action-Value function approximation with continuous wealth and discretized action could be converged to the analytical solution via semi-gradient SARSA.

Besides that, we change the global parameters such as ALPHA, BETA, Probability that the risky asset becomes α, and the maximum wealth to test the robustness of approximation learned by semi-gradient SARSA.

Multiple Q Models:

From the analytical solution, we know that Q_function depends on T, that is, for each period,$Q_t(w_t, x_t)$ correspond to different b_t and c_t Thus, we try on training multiple Action-Value function simultaneously to discover whether such model is useful.

We test the model in learning 2 Action-Value Functions simultaneously and expand the model to 10 Action-Value Functions. Similarly, we we change the global parameters to test the robustness.