We are given wealth $W_0$ at time 0. At each of discrete time steps labeled $t = 0, 1, \ldots, T$, we are allowed to allocate the current wealth $W_t$ in a risky asset and a riskless asset in an unconstrained, frictionless manner. The risky asset yields a random rate of return $\sim N(\mu, \sigma^2)$ over each single time step. The riskless asset yields a rate of return denoted by $r$ over each single time step.

Our goal is to maximize the Utility of Wealth at the final time step $t = T$ by dynamically allocating $x_t$ in the risky asset and the remaining $W_t - x_t$ in the riskless asset for each $t = 0, 1, \ldots, T - 1$ (assume no transaction costs and no restrictions on going long or short in either asset). Assume the single-time-step discount factor is $\gamma$ and the Utility of Wealth at the final time step $t = T$ is $U(W_T) = -\frac{e^{-aW_T}}{a}$ for some fixed $a > 0$.

- Formulate this problem as a Continuous States, Continuous Actions MDP by specifying it’s State Transitions, Rewards and Discount Factor. The problem then is to find the Optimal Policy.

State will be represented as $(t, W_t)$. Assume our decision (Action) at any time step $t$ is given by the quantity of investment in the risky asset at time step $t = 0, 1, \ldots, T - 1$ and is denoted by $x_t$ (hence, quantity of investment in the riskless asset at time $t$ will be $W_t - x_t$). We denote the policy as $\pi$, so $\pi((t, W_t)) = x_t$. Denote the random variable for the return of the risky asset as $R \sim N(\mu, \sigma^2)$ and the excess return of the risky asset (over riskless return $r$) as $S = R - r$. So,

$$W_{t+1} = x_t(1 + R) + (W_t - x_t)(1 + r) = x_tS + W_t(1 + r)$$

The Reward is always 0 for all $t = 0, 1, \ldots, T - 1$ and the Reward at the terminal time step $t = T$ is $U(W_T) = -\frac{e^{-aW_T}}{a}$. The MDP discount factor is $\gamma$.

- As always, we strive to find the Optimal Value Function. The first step in determining the Optimal Value Function is to write the Bellman Optimality Equation.

We denote the Value Function for a given policy as:

$$V^\pi(t, W_t) = E_\pi[\gamma^{T-t} \cdot U(W_T) | (t, W_t)] = E_\pi[-\gamma^{T-t} \cdot \frac{e^{-aW_T}}{a} | (t, W_t)]$$
We denote the Optimal Value Function as:

\[ V^*(t, W_t) = \max_{\pi} V^\pi(t, W_t) = \max_{\pi} E_{\pi}[\gamma^{T-t} \cdot e^{-aW_T}/a | (t, W_t)] \]

The Bellman Optimality Equation is:

\[ V^*(t, W_t) = \max_{x_t} \{E_{R \sim N(\mu, \sigma^2)}[-\gamma \cdot V^*(t+1, W_{t+1})]\} \]

- Assume the functional form for the Optimal Value Function is 
  
  \[ -b_t e^{-c_t W_t} \]

  where \( b_t, c_t \) are unknown functions of only \( t \). Express the Bellman Optimality Equation using this functional form for the Optimal Value Function.

\[ V^*(t, W_t) = \max_{x_t} \{E_{R \sim N(\mu, \sigma^2)}[-\gamma \cdot b_{t+1} e^{-c_{t+1} (x_t S_t + W_t (1+r))}]\} \]

- Since the right-hand-side of the Bellman Optimality Equation involves a max over \( x_t \), we can say that the partial derivative of the term inside the max with respect to \( x_t \) is 0. This enables us to write the Optimal Allocation \( x_t^* \) in terms of \( c_{t+1} \).

\[ -c_{t+1} (\mu - r) + \sigma^2 c_{t+1}^2 x_t^* = 0 \implies x_t^* = \frac{\mu - r}{\sigma^2 c_{t+1}} \]

- Substituting this maximizing \( x_t^* \) in the Bellman Optimality Equation enables us to express \( b_t \) and \( c_t \) as recursive equations in terms of \( b_{t+1} \) and \( c_{t+1} \) respectively.

Plugging in \( x_t^* \) in the above equation for \( V^*(t, W_t) \) gives:

\[ V^*(t, W_t) = -\gamma \cdot b_{t+1} e^{-c_{t+1} W_t (1+r) - (\mu-r)^2/2\sigma^2} \]

But since

\[ V^*(t, W_t) = -b_t e^{-c_t W_t} \]

we can write the following recursive equations for \( b_t \) and \( c_t \).

\[ b_t = \gamma \cdot b_{t+1} e^{-(\mu-r)^2/2\sigma^2} \]
\[ c_t = c_{t+1} (1 + r) \]

- We know \( b_T \) and \( c_T \) from the knowledge of the MDP Reward at \( t = T \) (Utility of Terminal Wealth), which enables us to unroll the above recursions for \( b_t \) and \( c_t \).

Since \( V^*(T, W_T) = \frac{e^{-aW_T}}{a}, b_T = \frac{1}{a}, c_T = a \). Therefore, we can unroll the above recursions for \( b_t \) and \( c_t \).
\[ b_t = \gamma^{T-t} \cdot \frac{a}{a} \cdot e^{-\frac{(\mu - r)^2}{2\sigma^2} \cdot (T-t)} \]
\[ c_t = a \cdot (1 + r)^{T-t} \]

- Solving \( b_t \) and \( c_t \) yields the Optimal Policy and the Optimal Value Function.

\[ x_t^* = \frac{\mu - r}{\sigma^2 a (1 + r)^{T-t-1}} \]
\[ V^*(t, W_t) = -\gamma^{T-t} \cdot \frac{a}{a} \cdot e^{-\frac{(\mu - r)^2}{2\sigma^2} \cdot (T-t)} \cdot e^{-a \cdot (1+r)^{T-t-1} \cdot W_t} \]