UCLouvain

LINMA2222 - Stochastic Optimal Control & RL
Project - Part 1

Optimal Portfolio Strategy

Students
Lucas Ahou (35942200)
Aymeric Couplet (59482200)

Teachers
R. Jungers
G. Bianchin
G. Berger

October 10, 2025

Question 2.1

The expression of g_t can be seen as a composition of two terms:

$$g_t := 1000 \cdot \left(q_{t+1}p_{t+1} - q_t p_t - (q_{t+1} - q_t)\overline{p}_{t,t+1}\right)$$

The first term (in blue) represents the change in value between time t and $t+1$ of our holdings with respect to the market price.

The second term (in red) represents the cash flow from spending money on stocks or selling them at the execution price $\overline{p}_{t,t+1}$. If $q_{t+1} > q_t$, then we bought some stocks and, similarly, if $q_{t+1} < q_t$ then we sold some stocks.

We can thus conclude that g_t represents our (gross) profit/loss between t and $t+1$.

Question 2.2

The price at the next time is equal to the last time plus 4 factors : $p_{t+1} = p_t + z_t^a + z_t^u + \gamma^u u_t + \sigma^p \xi_t^p$

- $\boldsymbol{z_t^a}$: This state variable represents the influence of random factors in the evolution of price (e.g. political situation, climate, etc.). It reverts gradually to zero with persistence controlled by ω^a and id driven by normally distributed changes of variance $(\omega^a \sigma^a)^2$. This implies that small values of ω^a create a slow decay of the random part ξ_t^a.
- $\boldsymbol{z_t^u}$: This state variable is deterministic and represents the temporary effect of the input u_t on the price. It takes the same form as z_t^a with a coefficient ω^u that controls the persistence and a weight β^u to scale the impact of the input. This simulates the impact of your purchase on the market price.
- $\boldsymbol{\gamma^u u_t}$: This term represents the immediate effect of u_t on the next price p_{t+1}. Unlike z_t^u, this term has an effect only on the next time step.
- $\boldsymbol{\sigma^p \xi_t^p}$: This term adds random and immediate fluctuations of the price to account for other phenomena that are not taken into account in the model.

All of this components make sense for an asset's price over time and we have all the major components for a reasonable representation of it. It takes into account external market factors as well as our trade decisions and have a decaying impact and an immediate one.

- $\boldsymbol{\overline{p}_{t,t+1}}$: Finally, the execution price is a weighted average of the price p_t and p_{t+1} using a parameter θ that interpolates both values.

Question 2.3

Let $y_t = (x_t^\top, u_t, \xi_t^\top)^\top = (q_t, z_t^a, z_t^u, u_t, \xi_t^a, \xi_t^p)^\top$. We take the expression of g_t :

$$\begin{aligned}
\frac{g_t}{1000} &= q_{t+1}p_{t+1} - q_t p_t - (q_{t+1} - q_t)\overline{p}_{t,t+1} \\
&= (q_t + u_t)p_{t+1} - q_t p_t - u_t \overline{p}_{t,t+1} \\
&= q_t(p_{t+1} - p_t) + u_t(p_{t+1} - \overline{p}_{t,t+1}) \\
&= q_t(z_t^a + z_t^u + \gamma^u u_t + \sigma^p \xi_t^p) + \theta u_t(p_{t+1} - p_t - \gamma^u u_t) \\
&= q_t(z_t^a + z_t^u + \gamma^u u_t + \sigma^p \xi_t^p) + \theta u_t(z_t^a + z_t^u + \sigma^p \xi_t^p)
\end{aligned}$$

We now have an expression that depends only on the variable y_t and can thus be put in a vector form :

$$\frac{g_t}{1000} = y_t^\top \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{\gamma^u}{2} & 0 & \frac{\sigma^p}{2} \\ \frac{1}{2} & 0 & 0 & \frac{\theta}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{\theta}{2} & 0 & 0 \\ \frac{\gamma^u}{2} & \frac{\theta}{2} & \frac{\theta}{2} & 0 & 0 & \frac{\theta\sigma^p}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{\sigma^p}{2} & 0 & 0 & \frac{\theta\sigma^p}{2} & 0 & 0 \end{bmatrix} y_t$$

$$\implies g_t = \frac{1}{2} y_t^\top H y_t$$

where:

$$H = 1000 \cdot \begin{bmatrix} 0 & 1 & 1 & \gamma^u & 0 & \sigma^p \\ 1 & 0 & 0 & \theta & 0 & 0 \\ 1 & 0 & 0 & \theta & 0 & 0 \\ \gamma^u & \theta & \theta & 0 & 0 & \theta\sigma^p \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma^p & 0 & 0 & \theta\sigma^p & 0 & 0 \end{bmatrix}$$

Question 2.4

As the statement suggests, let's consider that g_t is a random variable. In particular, let's assume it is a uniformly distributed RV. To be zero-mean and of variance σ^2, g_t must follow this distribution:

$$g_t \sim \mathcal{U}\left(-\sqrt{3}\sigma, \sqrt{3}\sigma\right)$$

$$\Rightarrow \varphi_{g_t}(g) = \begin{cases} \frac{1}{2\sqrt{3}\sigma} & \text{if } x \in \left[-\sqrt{3}\sigma, \sqrt{3}\sigma\right] \\ 0 & \text{otherwise} \end{cases}$$

We can now compute the expectation of the utility function with respect to g_t:

$$
\begin{aligned}
\mathbb{E}[c(g_t)] &= \int_{\mathbb{R}} c(g)\varphi_{g_t}(g)\,\mathrm{d}g \\
&= \int_{-\sqrt{3}\sigma}^{\sqrt{3}\sigma} \max\left(g - \frac{1}{2}g^2, 1 - \exp(-g)\right) \frac{1}{2\sqrt{3}\sigma}\,\mathrm{d}g \\
&= \frac{1}{2\sqrt{3}\sigma} \int_{-\sqrt{3}\sigma}^{\sqrt{3}\sigma} \max\left(g - \frac{1}{2}g^2, 1 - \exp(-g)\right)\mathrm{d}g \\
&:= F(\sigma)
\end{aligned}
$$

On Figure 1, we plotted $F(\sigma)$ for positive values of σ. We observe that the function is indeed decreasing as σ increases.

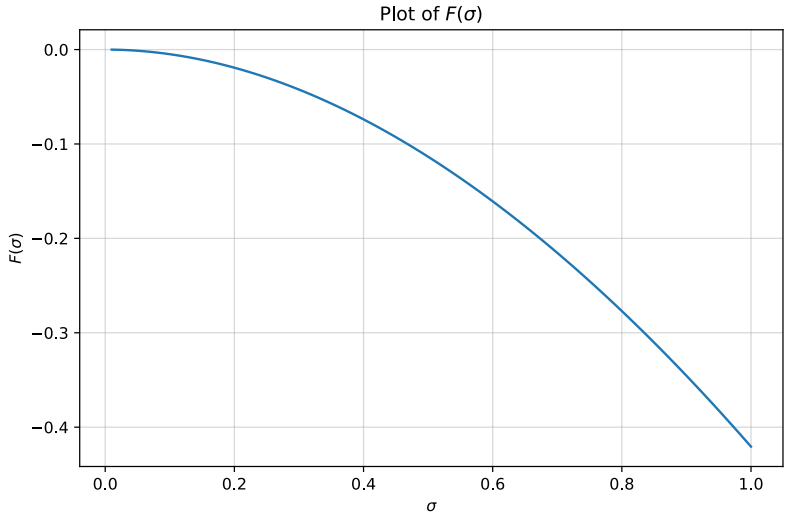


Figure 1: Plot of $F(\sigma)$ for $\sigma > 0$

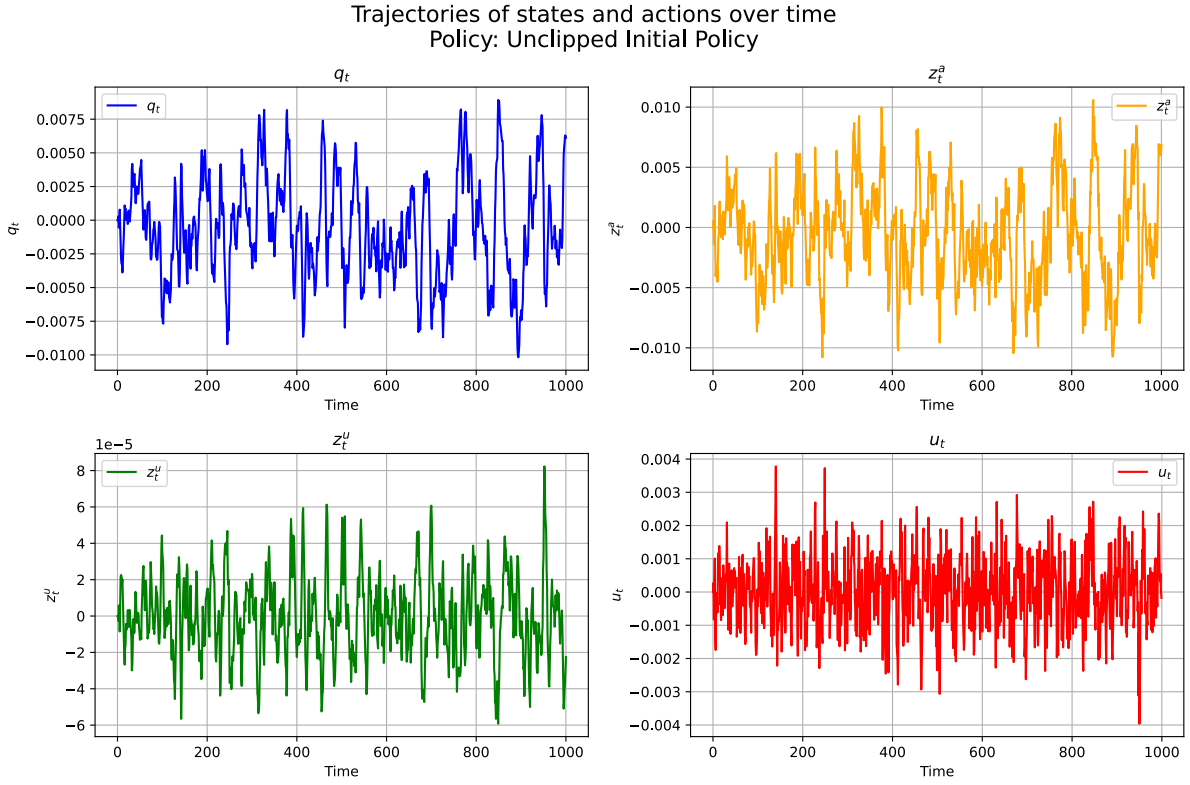
Question 3.1

Let us now simulate the evolution of the dynamical model with the policy π_{cl} defined as:

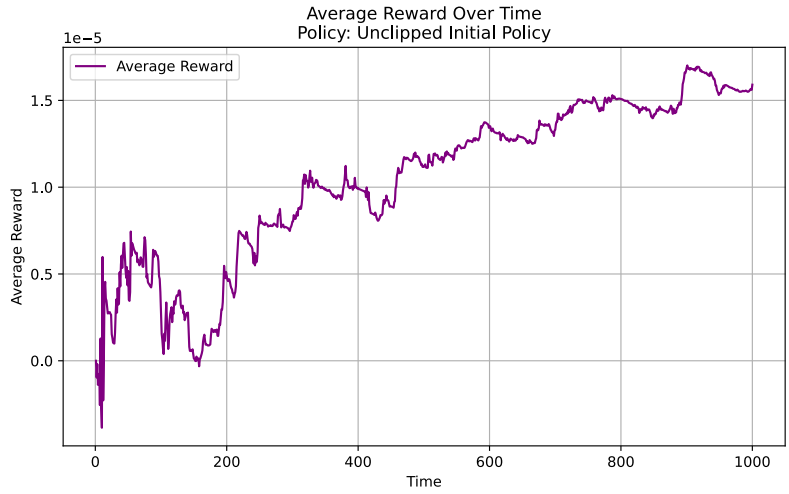
$$\pi_{\text{cl}}(x_t) = K_{\text{cl}} x_t = -0.5 q_t + 0.5 z_t^a + 0.5 z_t^u$$

with a initial condition $x_0 = (0, 0, 0)$ and for time $t = 1, ..., 1000$.

Here is the plots of the three state variables and the input u_t.



We can observe that the x_t and u_t both seem to be centered around 0. This means that there is as many moments where we buy than moments where we sell assets. Of course, this is not realistic as it results in negative quantities of stocks in our portfolio. This aspect is taken into account in the next strategy where we constrain $q_t \geq 0$.



Here is the average reward between 0 and t for each time t.
We observe that it ends up to a slightly positive constant ($\approx 1.5 \times 10^{-5}$). This is an indicator that this is a winning policy.

Question 3.2

Let us now run a simulation of the model in the same way but for $N = 100$ trajectories.
On Figure 4, we plotted the empirical distribution of the state variables. We can now clearly see that they are all centered around 0. Concerning the variance, we see that q_t and z_t^a seem to have approximately the same variance. Overall, they all seem to follow a normal distribution. z_t^u, however, has a very small variance. This is surely due to the fact that it does not, directly, depends on a random variable.

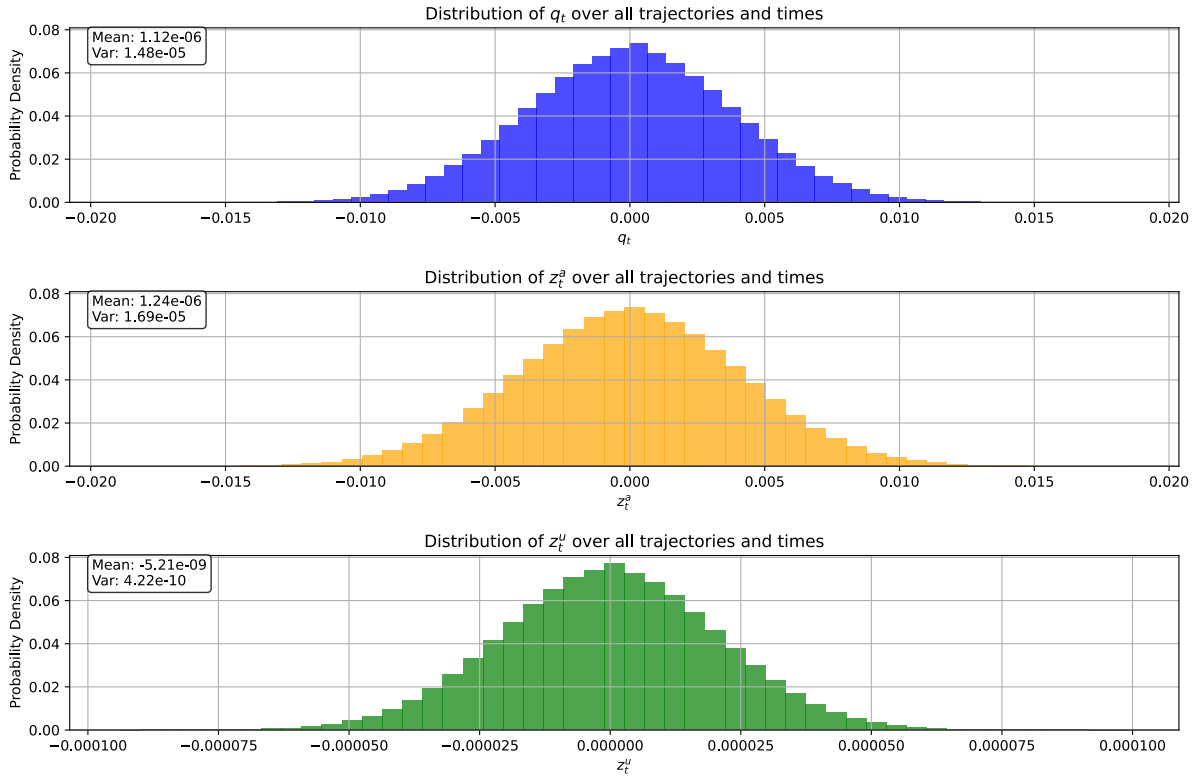


Figure 4: Empirical distribution of the state variables

On Figure 5, we plotted all the average rewards trajectories and their average (in purple) for $N = 1000$ trajectories and the distribution of the average reward for every trajectory.
We can see at the start, the reward have a high variance and can still be negative but as time goes the average reward tends to a positive $\approx 1.58 \times 10^{-5}$. This means this strategy is profitable in the long term. If we look at the empirical distribution, we see that after 1000 time steps there is practically no chance of loosing.

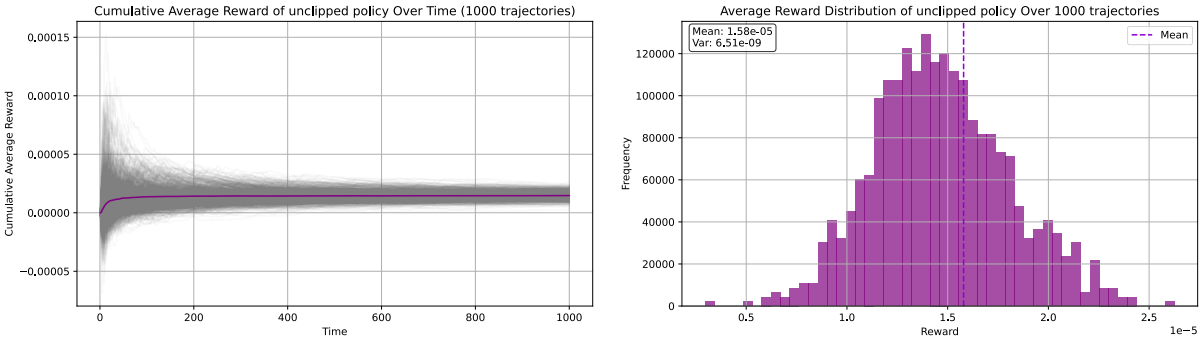


Figure 5: Plot of the average rewards for every trajectory

Question 3.3

The previous policy was not realistic because it could result in negative quantities of assets in our portfolio. To fix this, we consider the following policy: $\pi_{\text{pcl}}(x_t) = \max(-q_t, \min(\pi_{\text{cl}}(x_t)))$ which returns the previously defined policy if it does not break the constraint $q_t \geq 0$, otherwise it clips the returned value between $[-q_t, 1 - q_t]$ which prevents the input to change the value of our portfolio to exceed the range $[0, 1]$.

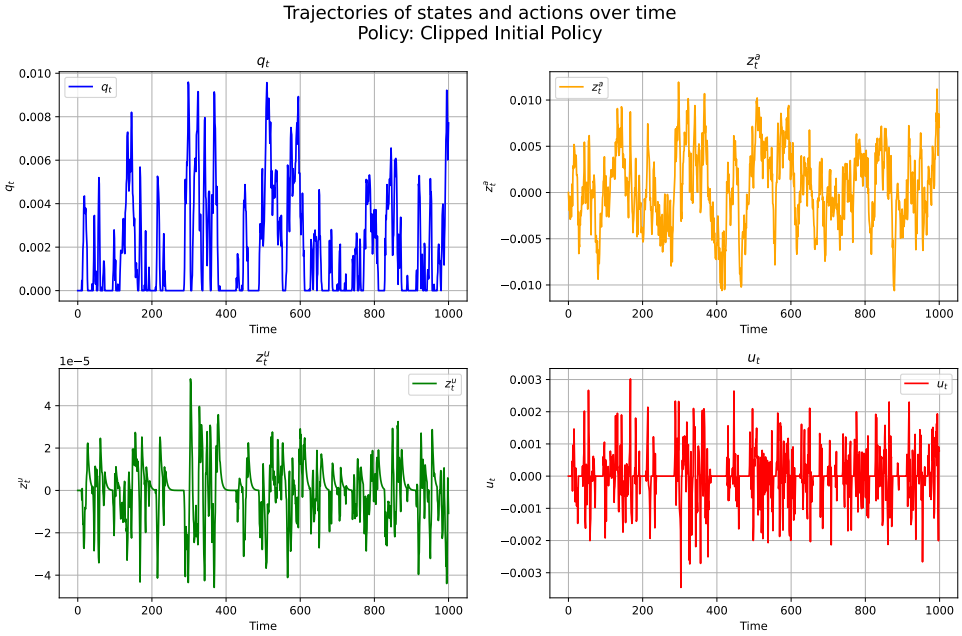
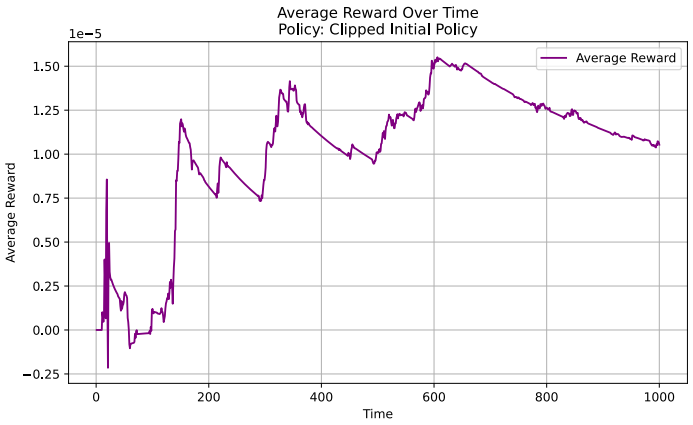


Figure 6: Plots of state variables and input u_t for the clipped policy

On Figure 6, we first notice that q_t indeed satisfies the constrained range. We can also notice some plateaux for some regions for q_t and u_t. Those plateaux represent moments where we previously wanted to reach a negative quantity q_t in our portfolio but the policy won't allow it. It thus result in a clipped value to 0. For those regions, we also observe that z_t^u decays to 0 because the input is null.



What is more interesting is that we observe multiple drops in the average reward over time and that those drops correspond exactly to the time the plateaux occur. An explanation for this would be that the unclipped policy would have been able to leverage the loss – or even win – but the constraint prevents those actions from being made which results in unwanted choice from our policy. Of course, as a consequence of this, we observe that the long term average reward seems to tend to a lower value.

Question 3.4

We ran the simulation again for $N = 100$ trajectories with the clipped policy. On Figure 8, we clearly see that q_t is always non-negative. Consequently, its mean is greater than for the previous policy. It also results in the probability for q_t to be equal to 0 to be very high, as all previously negative values are now projected to 0. The same goes for z_t^u where we can clearly see the previously mentioned decays due to the input u_t being null.

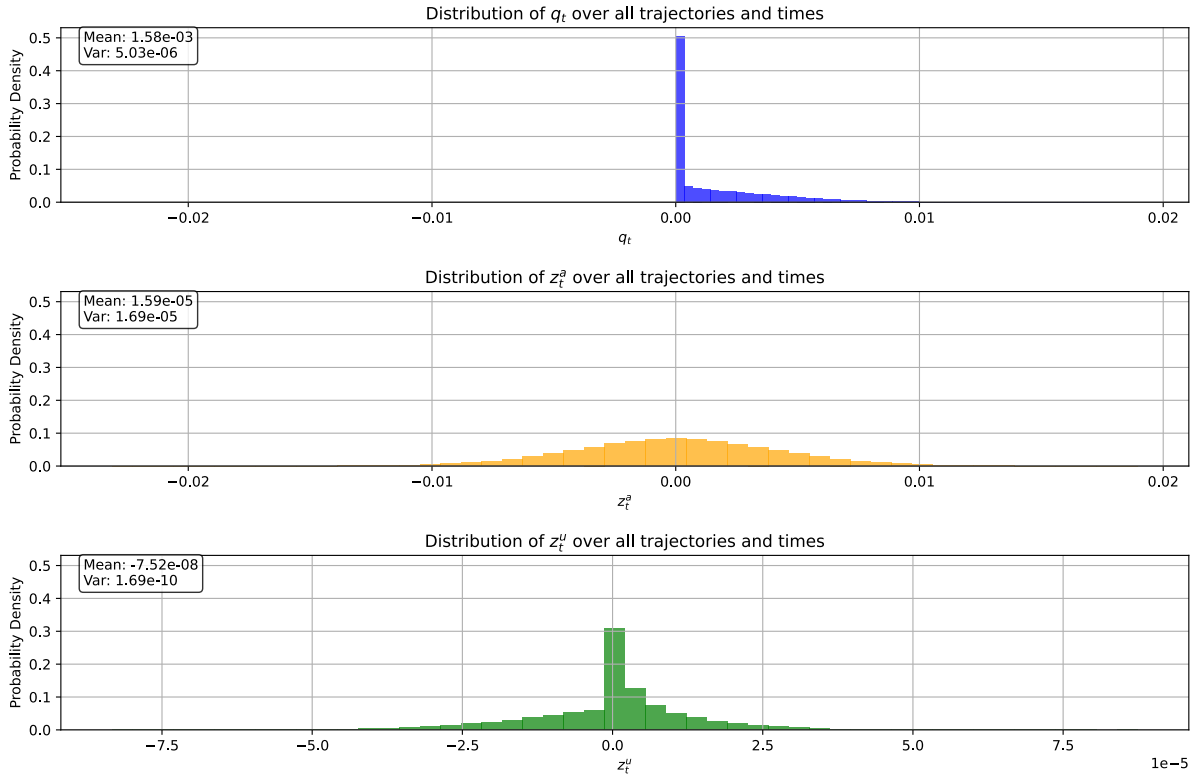


Figure 8: Empirical distribution of the state variables

On Figure 9, we finally plotted the reward distribution like before. The variance at steady-state seems to be of the same magnitude as for the previous policy. However, as mentioned in the previous section, we observe that the infinite horizon average reward seems to have a smaller mean. This is because, while this policy is more realistic, it does not permit to make "illegal" decisions which allowed us to make more profit.

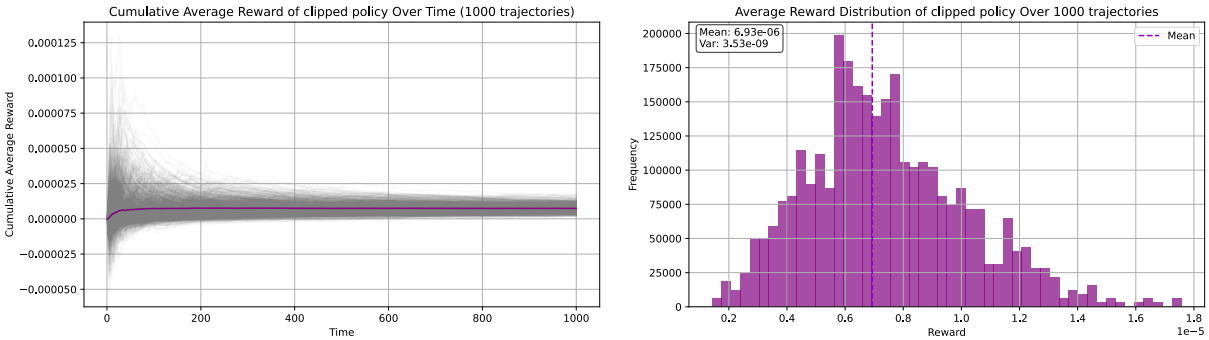


Figure 9: Plot of the average rewards and the distribution at the final time step

Question 3.5

To find a better policy, we will to optimize the linear policy :

$$\pi_l^*(x_t) = \max\Big(q_t, \min\Big(1 - q_t, p_3[q_t, z_t^a, z_t^u]^\top\Big)\Big)$$

where $p_3 = [\alpha, \beta, \gamma]$ is the array of 3 parameters. Of course, we clip the value of the output to satisfy the constraints.

We optimized the reward at final time step $T = 1000$ averaged on $N = 1000$ simulations with CMA-ES solver from the `cma` library. This solver is made for stochastic function as opposed to a solver like `SciPy` and works way better.

The solver ran tests 20.000 times at $N = T = 1000$ for about 20 minutes and found $p_3 = [-0.0529, 37.2788, 29.9648]$

The linear optimal policy is a bit limited. That's why we added quadratic relations so it can only do better :

$$\pi_q^*(x_t) = \max\Big(q_t, \min\Big(1 - q_t, p_{10}\left[q_t, z_t^a, z_t^u, (q_t)^2, (z_t^a)^2, (z_t^u)^2, q_t z_t^a, z_t^a z_t^u, z_t^u q_t, 1\right]^\top\Big)\Big)$$

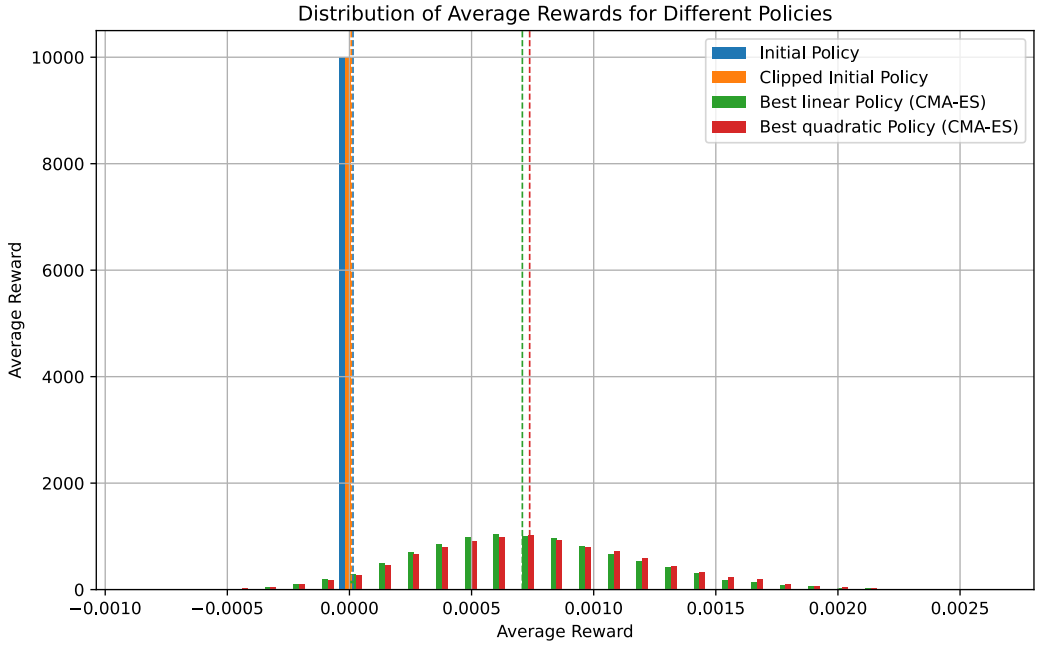
where p_{10}, the 10 parameters to optimize this quadratic policy.

$p_{10} = [-0.1674, 25.6970, 7.1492, 0.0350, 17.5269, -4.0057, 9.3930, 24.3607, -0.5704, 0.0603]$

Comparison

To compare all the policies infinite-horizon average reward, we plotted their distributions. We can clearly see how our policies performs way better than the previous two, even though there is a larger variance in the average reward.

The quadratic policy performed marginally better than the linear policy. Let's analyze this quadratic policy.



Question 3.6

Again, let us plot the states variables, the input, as well as the average reward for a single trajectory.
On Figure 11, we first notice that q_t satisfies the constraint at all time steps. We also notice that q_t reaches extreme values more often than for the other policies. This seems pretty logic since we logically want to buy all the stocks available if there is a good opportunity, and we want to sell all of our stocks if we predict that the price will go down.

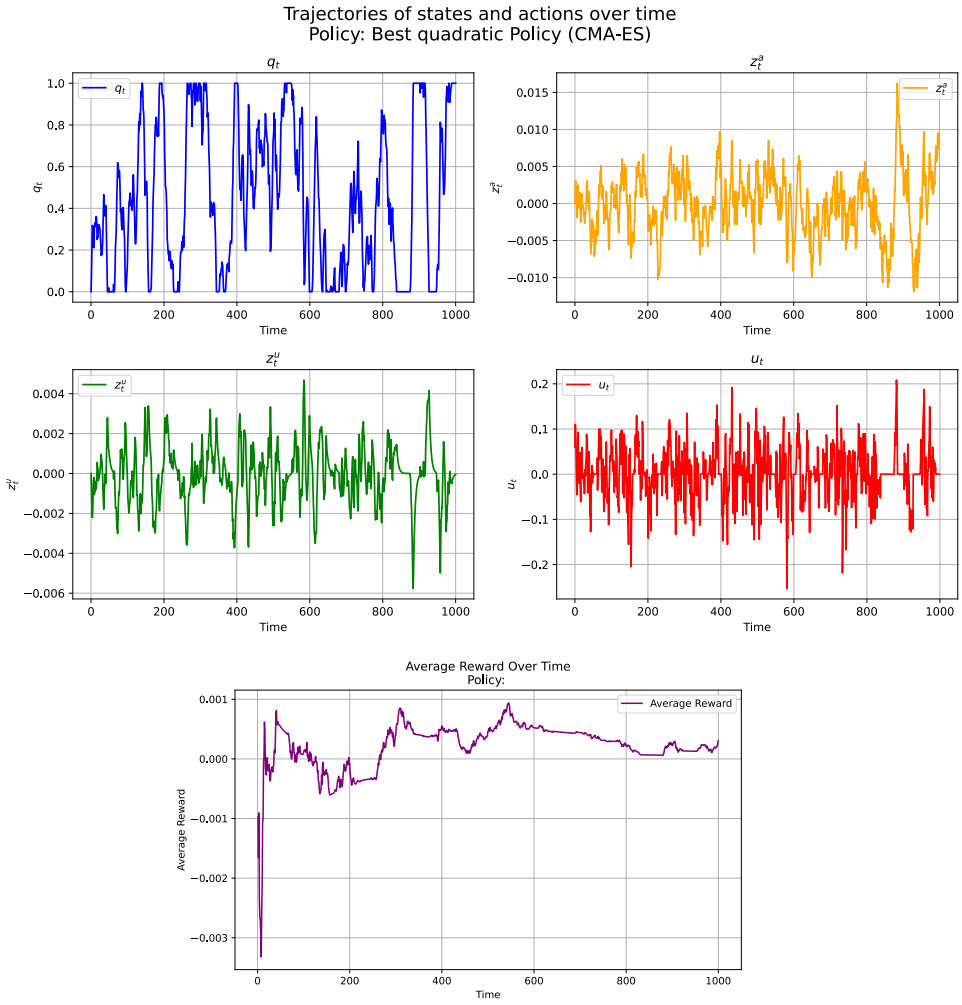


Figure 11: Plots of state variables, input u_t and average reward for our policy

Question 3.7

Now, let's run 100 trajectories and take a look at the empirical distributions as before. On Figure 12, we clearly observe the phenomenon mentioned above where the values of q_t are mainly 0 or 1.

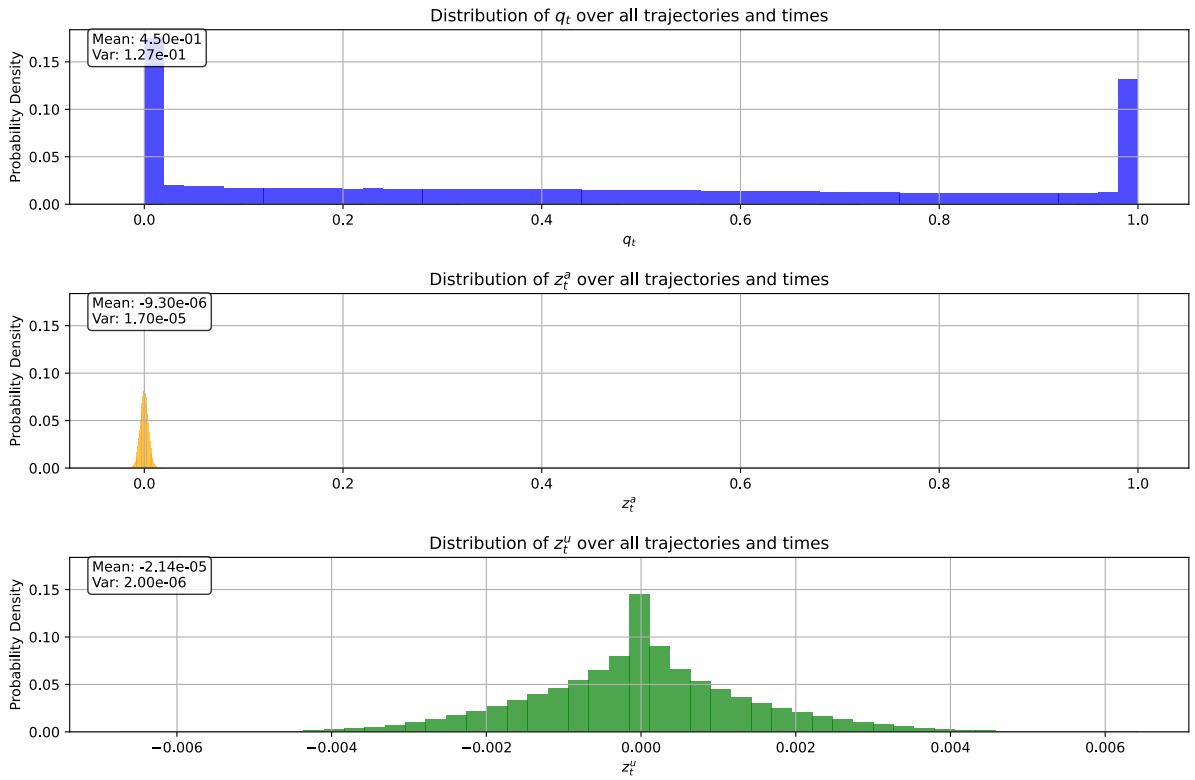


Figure 12: Empirical distribution of the state variables

What is more interesting is that we observe a way better infinite-horizon average reward for this policy ($\approx 7 \cdot 10^{-4}$). It is 10 times higher the first, non clipped policy.

The drawback is that there is now a higher probability to get a negative reward after $T = 1000$ time steps.

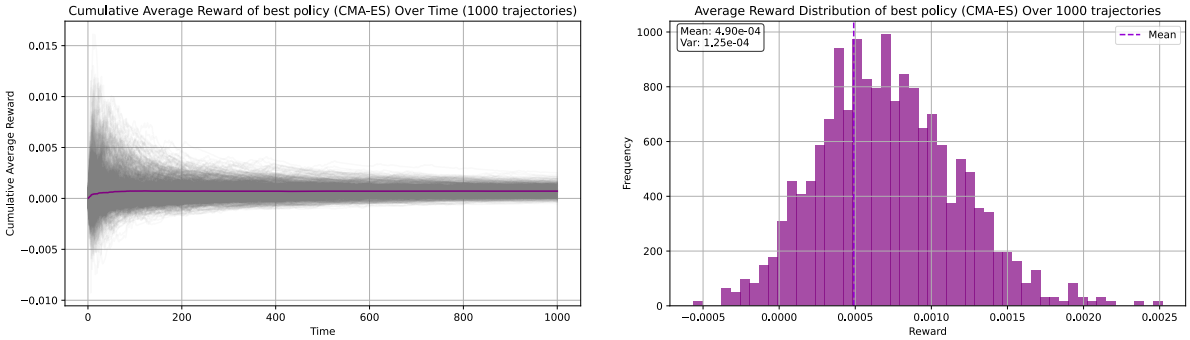


Figure 13: Plot of the average rewards and the distribution at the final time step