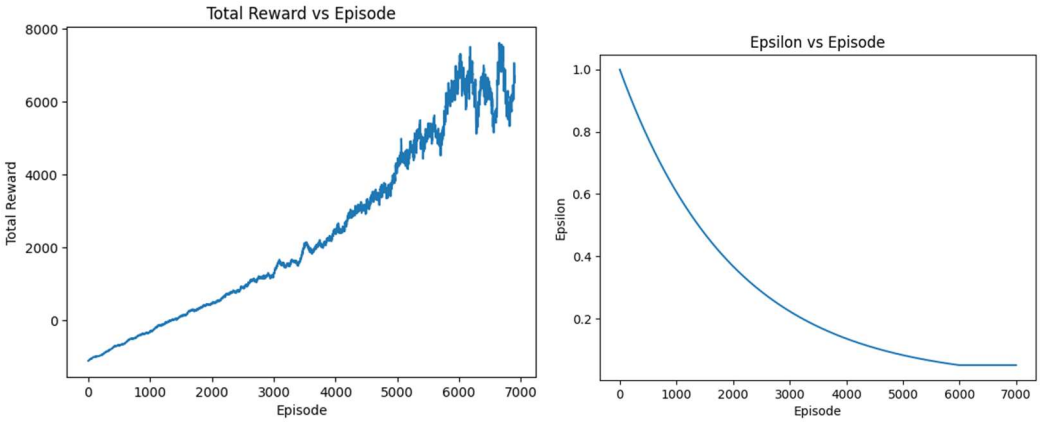
Part 3 [Total: 30 points] - Solve Stock Trading Environment

Show and discuss the results after applying the Q-learning algorithm to solve the stock trading problem. Plots should include epsilon decay and total reward per episode.



Total Rewards per episode:

- The provided plot shows an increasing trend in total rewards, which implies that the Q-learning agent is acquiring a useful trading policy.
- The reward curve suggests a trend of better decision making with more regular high rewards in future episodes.
- Somewhere around episode 5000, one sees a definitive sharp rise in rewards, which may indicate that the agent is adopting a better strategy.

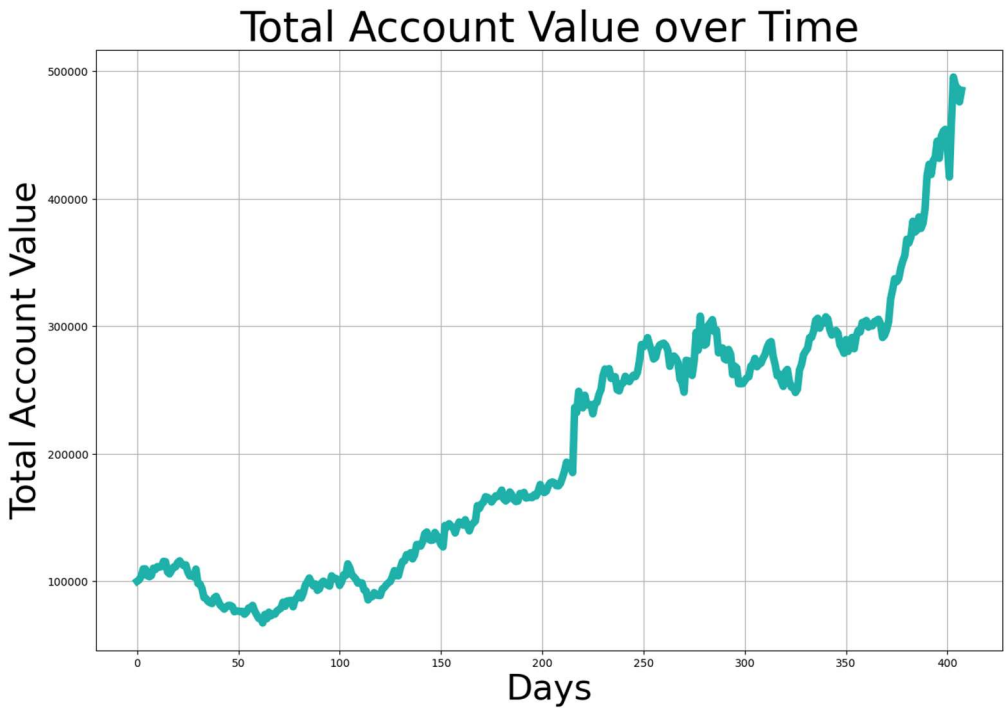
Epsilon Decay

- Q learning would start with a high exploration rate and then decrease it, skewed towards exploiting learned strategies.
- If the decay of epsilon is well-tuned, the agent learns sufficiently in early episodes and converges to a best trading strategy subsequently.

Insights:

- Erratic behavior in the final episodes may suggest the need for further tuning or testing on unseen samples.

Provide the evaluation results. Evaluate your trained agent's performance (you will have to set the train parameter to False), by only choosing greedy actions from the learnt policy. The plot should include the agent's account value over time. Code for generating this plot is provided in the environment's render method. Just call environment.render after termination.



Explanation of plot:

- **The graph shows account value in terms of time when agent is probed with greedy policy (only the best-trained actions are chosen).**
- **Firstly, the account balance varies minimally, indicating market volatility and initial decision making.**
- **At day 200, a sudden increase in account value shows that agents have learned effective trading strategies and is making good choices.**
- **The explosive growth after 300 days demonstrates excellent performance, as the agent is exploiting the trends in market.**
- **End-of-account value is much greater than the initial values, which proves that learned policy was successful in increasing portfolio**