Reinforcement Learning Model Comparison for Trade Scheduling

Prem Shenoy

October 26, 2024

Introduction

This report covers the performance comparison of two reinforcement learning algorithms for optimizing trade scheduling: Soft Actor-Critic (SAC) and Proximal Policy Optimization (PPO). Both models were evaluated on minimizing transaction costs while selling a fixed quantity of shares. The results of the experiments and the training dynamics are discussed.

Proximal Policy Optimization (PPO)

PPO is another state-of-the-art reinforcement learning algorithm that optimizes a clipped objective function. It improves training stability by avoiding large updates that could destabilize learning.

Training Dynamics of PPO

The PPO model was trained over multiple episodes. Key statistics from the training process are as follows:

- **Training time per iteration**: Ranged between 1 and 45 seconds.
- **Total iterations**: 25
- **Average FPS**: Approximately 1,120 FPS.
- **Timesteps per iteration**: 2,048.
- **Entropy loss**: Started at -1.42 and remained constant throughout training, encouraging exploration.
- **Value loss**: The final value loss was 9.8×10^9.
- **Policy gradient loss**: The final policy gradient loss was approximately -0.000846.
- **Explained variance**: Remained low throughout training, with a final value of approximately 0.000214.

Key learning parameters included:

- **Learning rate**: 0.0003.
- **Clip range**: 0.2.
- **Updates per iteration**: 90–230 updates.

Results and Analysis

The PPO model's performance was compared to the SAC model and traditional benchmarks.

Strengths of PPO

- **Stable Policy Updates**: PPO maintained steady policy updates by limiting large changes, evidenced by the low policy gradient loss.
- **Exploration vs Exploitation Balance**: The entropy loss of -1.42 ensured that the model maintained a balance between exploration and exploitation.
- **Efficiency in Training**: PPO achieved efficient training, managing over 1,120 frames per second while keeping updates per iteration within 90–230.

Weaknesses of PPO

- **Low Explained Variance**: PPO's explained variance remained close to zero, indicating that it struggled to model the variance in future returns.
- **High Value Loss**: The value function showed inefficiencies, with the value loss reaching up to 9.8×10^9, implying difficulty in accurately estimating future rewards.

Comparison with SAC

The Soft Actor-Critic (SAC) algorithm is more suitable for environments requiring higher adaptability, such as volatile markets. Key differences between the two models include:

SAC's Advantages

- **Entropy Regularization**: SAC's entropy maximization encourages exploration, allowing the algorithm to adapt dynamically to unpredictable market conditions.
- **Superior in Volatile Markets**: SAC performs better in dynamic environments by exploring more widely, making it ideal for markets with high fluctuations.

SAC's Disadvantages

- **Higher Training Complexity**: SAC is more complex and slower to train due to its focus on entropy maximization, which increases its computational requirements compared to PPO.

Conclusion

This report highlights the use of both SAC and PPO for optimizing trade scheduling. SAC's entropy regularization allowed for better adaptability to fluctuating market conditions, while PPO maintained stable policy updates. In conclusion, PPO is a better choice for stable environments, while SAC excels in volatile and dynamic market conditions. The choice between the two algorithms depends on the specific requirements of the trading environment.