Reinforcement Learning : Term Paper

Date - 29 April 2025

Stable and Realistic Deep Reinforcement Learning in Portfolio Management.

Mentor - Dr. Shashi Shekhar Jha

Team members:
Shatrughna Chaurasia - 2024dsz0002
Peguda Akshitha - 2021csb1122

Introduction

Problem Statement:

Current Deep Reinforcement Learning (DRL) models in portfolio management face challenges in stability, risk-awareness, and market realism. They often rely on symmetric risk metrics like the Sharpe Ratio and are trained in oversimplified environments. This limits their practical deployment and interpretability. The problem is to develop a DRL framework that is stable, downside-risk aware, and better aligned with real market behavior.

Introduction

Motivation:

The paper "Deployability of Deep Reinforcement Learning in Portfolio Management" highlights two major concerns:

- No common baseline to compare various DRL agents.
- No clear method to verify if DRL agents truly learn optimal policies.

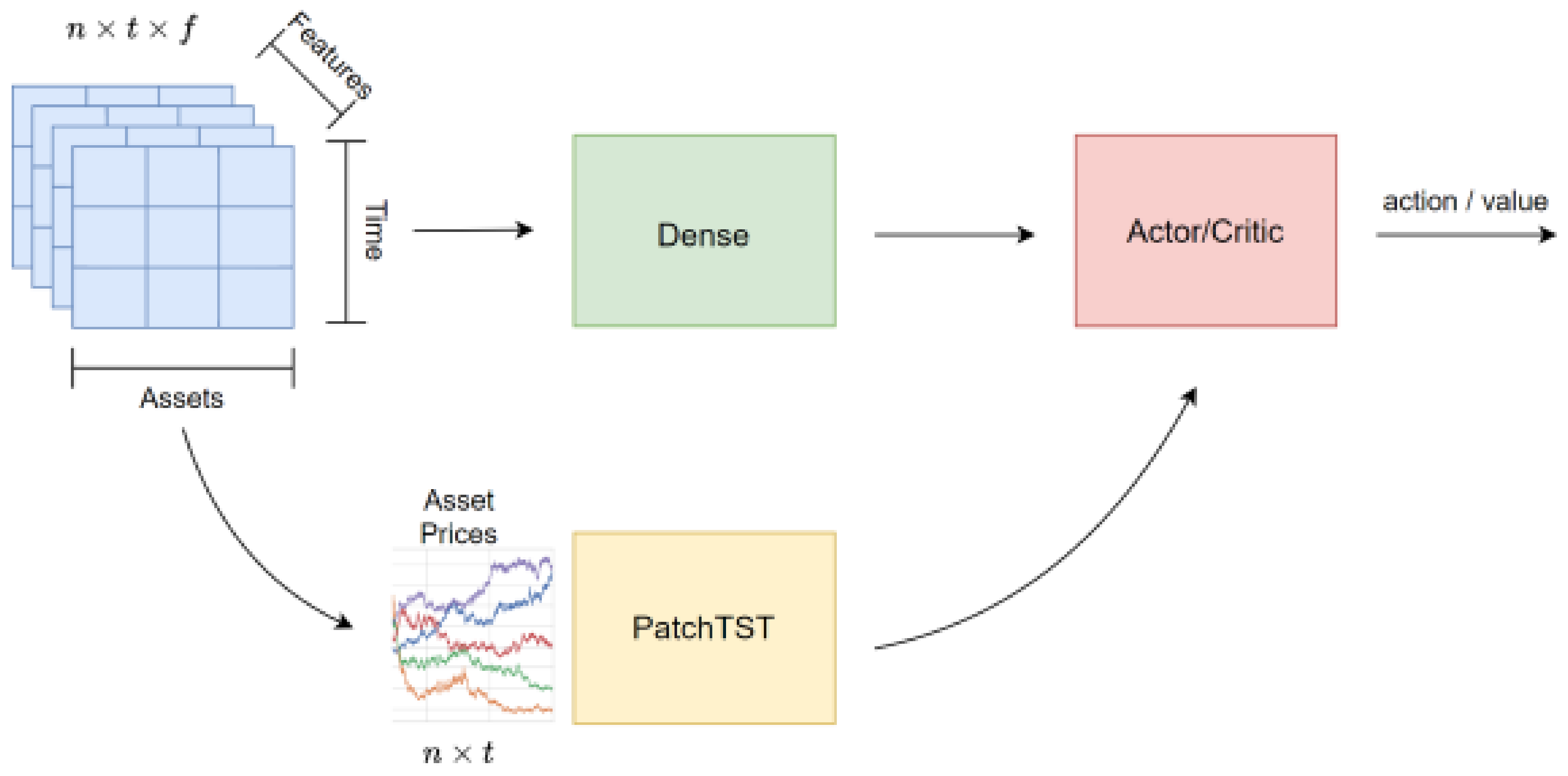
The problems in their solution :

- Existing DRL models (e.g., PPO + PatchTST, DA-RNN) show high variance
- Simulations use oversimplified market assumptions (e.g., BEKK)

Our goal is to enhance the stability and reliability of DRL-based portfolio management by addressing these limitations.

Proposed Approach

- Use of PPO-based DRL Agents
- Compare DRL with Max Sharpe (Tangency Portfolio)
- Create Synthetic Market using BEKK Model
- Design Feature Extractor Architectures
- Use Differential Sharpe Ratio (DSR) as Reward
- Evaluate Across Key Metrics



Architecture of the model

MDP

- **State** : At time t with time window l, the state is given by: $s_t = ((p_{t-l}, f_{t-l}, w_{t-l}), \ldots, (p_t, f_t, w_t)).$
- **Action**: The action at time t is the agent's attempt at reallocating the weights to rebalance decisions (portfolio weights for each asset)
  - Continuous values in [−1,1] or [0,1]
  - Positive = long, Negative = short
- **Transition**: Market dynamics simulated using BEKK model
  - Returns rt N(μ,Σ)rt N(μ,Σ)
  - Update prices using log returns: $\log \frac{p_{t+1}^{(i)}}{p_t^{(i)}} = r_t^{(i)}$
- **Reward**: Differential Sharpe Ratio (DSR) — risk-adjusted return using moving averages

$$\begin{aligned} &\underset{\mathbf{w}}{\text{maximize}} && \frac{\mu^T w - r_f}{(w^T \Sigma w)^{1/2}} \\ &\text{subject to} && \mathbf{1}^T w = 1 \\ &&& w_i \geq 0, \text{ if non-shorting} \end{aligned}$$

Proposed Novelty

- Introduce Sortino Ratio as the new reward function
- Replace PPO with TRPO for better stability
- Use GARCH(1,1) or Jump-Diffusion to simulate realistic markets

Novelty

Risk measure: Sortino ratio vs Sharpe ratio

Sharpe Ratio:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma}$$

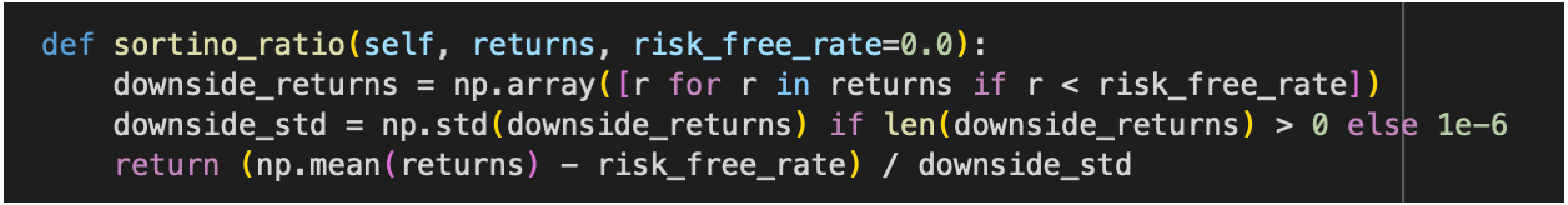
- Sharpe ratio penalizes both positive & negative variance

Sortino Ratio:

$$\text{Sortino Ratio} = \frac{R_p - R_f}{\sigma_d} \qquad \text{where} \qquad \sigma_d = \sqrt{\mathbb{E}\left[\{\min\left(\mu^T w - \mathbb{E}[R_p],\ 0\right)\}^2\right]}$$

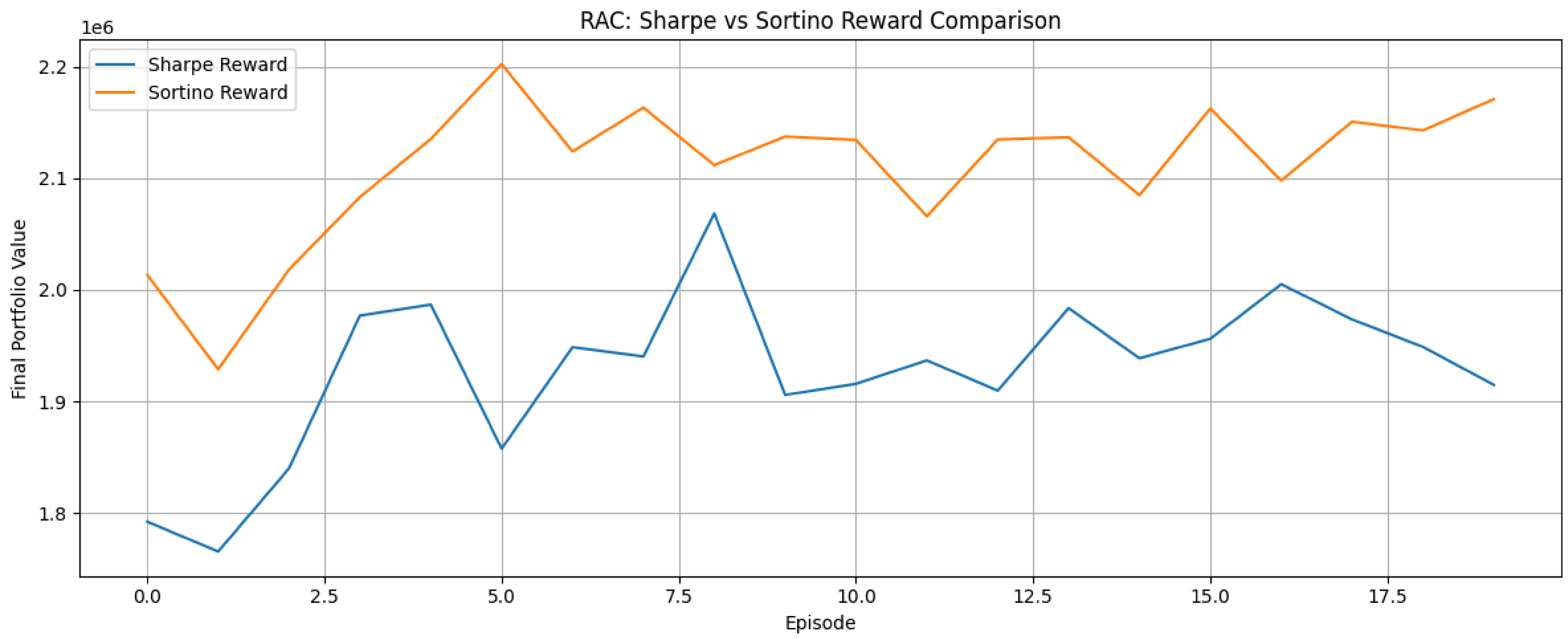
- Sortino ratio focuses only on downside risk (realistic risk)

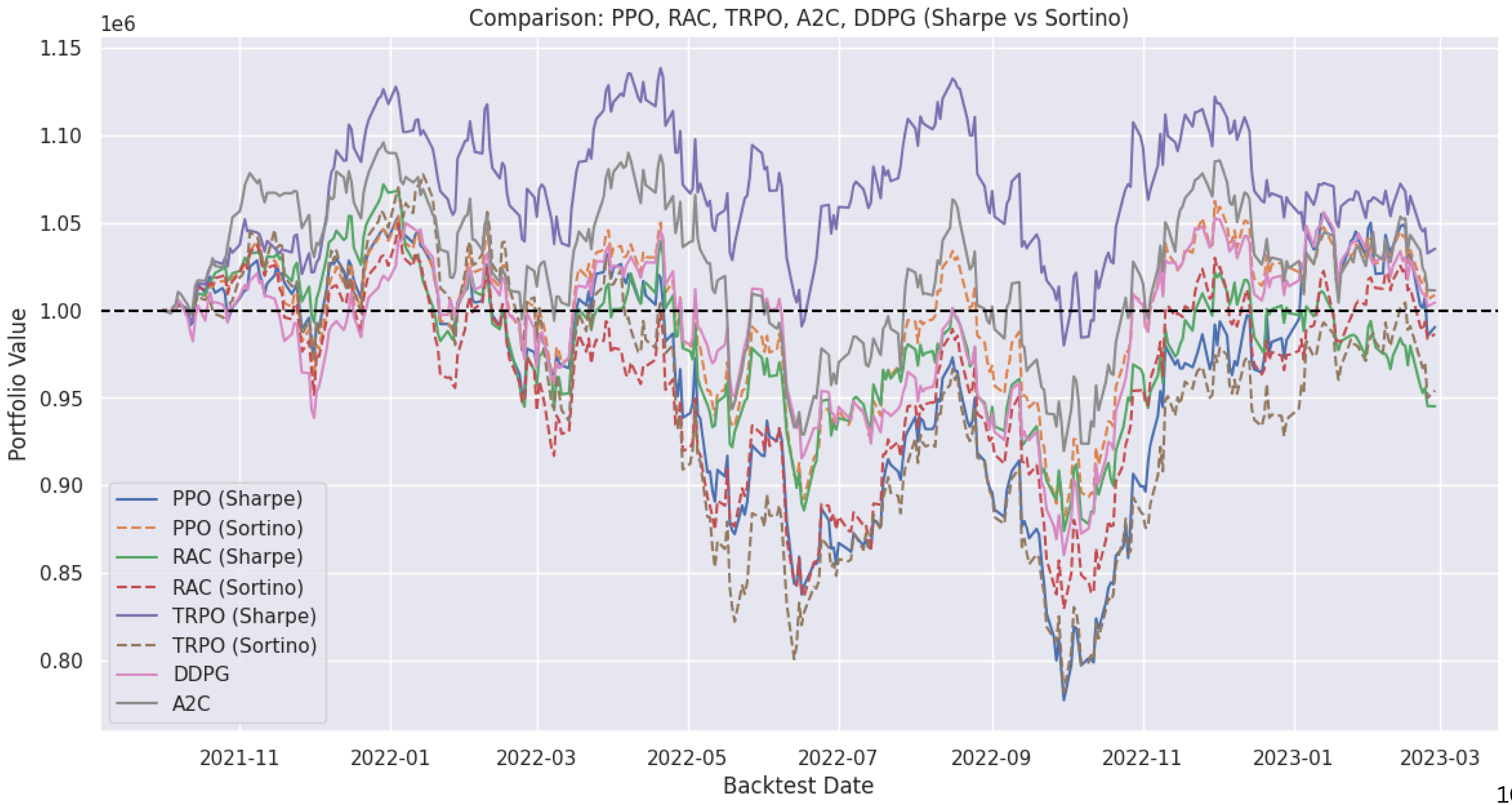
Sortino ratio



```python
def sortino_ratio(self, returns, risk_free_rate=0.0):
    downside_returns = np.array([r for r in returns if r < risk_free_rate])
    downside_std = np.std(downside_returns) if len(downside_returns) > 0 else 1e-6
    return (np.mean(returns) - risk_free_rate) / downside_std
```

Results


RAC: Sharpe vs Sortino Reward Comparison
1e6
Sharpe Reward
Sortino Reward
2.2
2.1
2.0
1.9
1.8
Final Portfolio Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Episode


Comparison: PPO, RAC, TRPO, A2C, DDPG (Sharpe vs Sortino)
1e6
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
Portfolio Value
PPO (Sharpe)
PPO (Sortino)
RAC (Sharpe)
RAC (Sortino)
TRPO (Sharpe)
TRPO (Sortino)
DDPG
A2C
2021-11
2022-01
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
2023-03
Backtest Date

MODEL	A2C	DDPG	PPO		TRPO		RAC	
			SORTINO	SHARPE	SORTINO	SHARPE	SORTINO	SHARPE
Annual Returns	0.812%	0.318%	0.622%	-0.682%	-3.289%	**2.494%**	-0.972%	-3.946%
Cumulative Returns	1.139%	0.446%	0.873%	-0.954%	-4.577%	**3.512%**	-1.359%	-5.483%
Annual Volatility	17.356%	**16.185%**	17.674%	19.897%	21.817%	16.796%	19.868%	17.219%
Sharpe Ratio	0.13	0.10	0.12	0.06	-0.04	**0.23**	0.05	-0.15
Max Drawdown	-16.066%	-18.059%	-16.642%	-26.189%	-27.413%	**-13.902%**	-20.96%	-18.464%
Daily Values at Risk	-2.177%	-2.033%	-2.218	-2.502%	**-2.753%**	-2.101%	-2.499%	-2.179%

Conclusion

- TRPO (Sharpe) delivers the best performance among all models, showing highest cumulative returns (3.51%), lowest drawdown (-13.9%), and stronger stability.
- TRPO is preferred because it offers better stability in learning due to trust-region constraints — ensuring more reliable policy updates even in volatile market conditions.
- If PPO is used, Sortino Ratio-based reward performs better than Sharpe-based reward.
- Sortino Ratio focuses only on downside risk, ignoring upside volatility. This leads to smarter risk control, favoring consistent returns over simply high returns.
- Sortino Ratio reward improves risk-sensitive learning, making models more defensive against market crashes.

Thank You