Heavy Weather:

PPO–Based Reinforcement Learning for Adaptive Portfolio Management Under Crisis Regimes

Hari Srikanth
University of California, Berkeley
Email: harinsrikanth@berkeley.edu

***Abstract*—Financial markets are riddled with regime shifts, tail risks, and non-stationary dynamics that render traditional portfolio strategies brittle under stress. In response, we present Heavy Weather, a multi-asset allocation framework powered by Proximal Policy Optimization (PPO). Our reinforcement learning agent is trained to navigate heterogeneous asset domains—equities, futures, and crypto—using a continuous action space and interpretable state representations (returns, volatility, VIX, portfolio weights). The reward function integrates realistic market frictions, drawdown penalties, and diversification incentives. Trained on historical daily data (2003–2019) and tested out-of-sample through 2023, Heavy Weather matches the returns of static heuristics like 60/40 and equal-weight portfolios in benign markets, but decisively outperforms them during crises such as the 2008 crash and 2020 pandemic drawdown. Notably, it delivers +4.8% in 2008 and flat performance in Q1 2020 where 60/40 lost 19%. Empirical evaluation across multiple metrics (Sharpe, Sortino, drawdown) and statistical tests (Welch's t-test, bootstrapping) confirms that PPO offers a viable, adaptive overlay to traditional allocation under regime uncertainty.**

I. Introduction

In April 2020, oil futures plunged below zero for the first time in history—an extreme event that broke many pricing models, invalidated assumptions about arbitrage constraints, and exposed the fragility of conventional hedging techniques. Such scenarios highlight a deeper truth: financial markets are highly non-stationary—characterised by abrupt structural breaks, heavy-tailed return distributions, and time-varying cross-asset correlations. Strategies that perform well under normal conditions, including momentum trading or volatility targeting, often fail catastrophically when the environment shifts into unfamiliar territory. This is further exemplified with challenging and volatile asset classes such as Cryptocurrencies, or stocks such as Tesla, which derives much of its value from arbitrary sentiment and market perception. Classical portfolio constructions predicated on covariance stationarity (e.g. MVO [1]) or static heuristics (e.g. 60/40) underperform during crises when diversification benefits evaporate [5]. Recent breakthroughs in *deep reinforcement learning* (RL) provide an adaptive alternative: agents learn directly from reward feedback, continually refining their policy to maximise long-horizon objectives under uncertainty [2].

In this paper I present **Heavy Weather**, a PPO-driven asset-allocation system expressly designed to handle *regime shifts*. Unlike value-based methods, PPO—as a policy-gradient algorithm—optimises a clipped surrogate objective, yielding stable updates even in noisy, non-stationary environments. Prior studies confirm PPO's efficacy in equities [4], commodities [6], and cryptocurrencies [7]. I extend this line of work by:

- unifying *three* heterogeneous asset buckets within a single continuous-action environment;
- engineering an interpretable state vector that mirrors human decision factors (returns, volatility, portfolio weights, VIX);
- incorporating realistic frictions—transaction costs, position constraints, crisis penalties—into the reward; and
- providing a rigorous statistical evaluation across multiple historical stress regimes.

Empirical results demonstrate that PPO dynamically rotates capital into safe-haven assets during turmoil and re-engages risk assets post-crash, yielding superior risk-adjusted returns over static baselines.

Code available at github.com/HariSrikanth/HeavyWeather

II. Background

A. Reinforcement Learning Paradigm

RL formalises sequential decision-making via a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$. At each time step t the agent observes state S_t, selects action $A_t \sim \pi_\theta(\cdot|S_t)$, receives reward R_t, and transitions to S_{t+1} (Fig. 1). The objective is to maximise expected discounted return $\mathbb{E}\big[\sum_t \gamma^t R_t\big]$. Two main algorithmic families exist:

- **Value-based** (e.g. Q-learning) estimate $Q^\pi(s, a)$ and derive greedy actions.
- **Policy-based** (e.g. Actor–Critic) update policy parameters directly via gradient ascent.

Actor–Critic combines both: an *actor* updates π_θ, while a *critic* approximates V^π. Natural Actor–Critic further pre-conditions gradients with the Fisher information matrix, expediting convergence [3]. PPO simplifies Trust-Region Policy Optimisation by clipping probability ratios, balancing exploration and stability [2].

B. RL in Quantitative Finance

Empirical evidence. Recent literature demonstrates RL's growing success in portfolio management:

- Wu *et al.* [4] showed a CNN-enhanced PPO that beat benchmarks on six-asset portfolios, noting that *monthly* rebalancing improved Sharpe.
- Santos *et al.* [5] compared PPO, SAC, DDPG, A2C, TD3 on US and Brazilian markets—RL portfolios realised ˜12% excess returns over 60/40 with similar volatility.

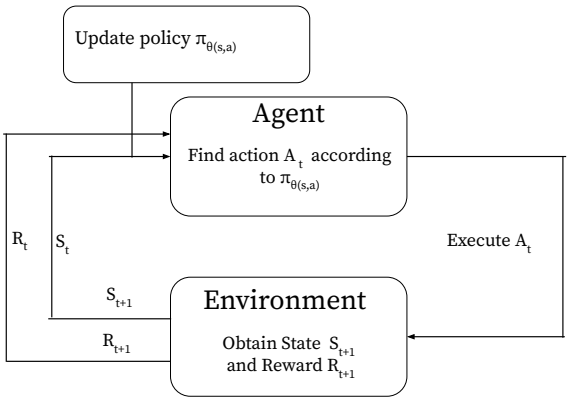

Fig. 1. Simplified RL loop: the policy π_θ chooses action A_t given state S_t, receives reward R_t, and updates via gradient ascent to maximise cumulative return.

TABLE I
ASSET CLASSES AND PROXIES

Domain	Ticker	Description
Traditional	SPY	S&P 500 ETF
	TLT	20Y US Treasury
	GLD	Gold ETF
Futures	USO	WTI Crude Oil ETF
	VIXY	VIX futures proxy
High-Vol	BTC-USD	Bitcoin spot
	ETH-USD	Ethereum spot
	ARKK	Innovation ETF

- Hanetho [6] achieved an 83% Sharpe uplift trading natural-gas futures using a volatility-aware PPO.
- Asgari [7] applied PPO to BTC/ETH, converting \$10k to \$14.85k over 66 days despite high crypto volatility.
- Jiang [8] pioneered deep RL for crypto using an Ensemble of Identical Independent Evaluators.

Open-source toolkits like **FinRL** [9] and TradeMaster standardise evaluation, emphasising careful risk management and hyper-parameter tuning.

III. PROBLEM FORMULATION

A. Asset Universe and Market Regimes

I construct three asset buckets (Table I) and evaluate over distinct regimes:
(i) Global Financial Crisis (2007–09), *(ii) COVID-19 crash* (2020-H1), and *(iii) Tech/Crypto bear* (2021–22). Regime selection follows guidance in Section II-B recommending stress-testing across equities, futures, and crypto extremes.

B. State Space $\mathcal{S}$

Each daily observation is a flattened vector:

$$S_t = [\underbrace{r_{t-1}^{(1:M)}, \cdots, r_{t-20}^{(1:M)}}_{\text{20-day returns}}, \underbrace{\text{RSI}_t^{(1:M)}, \text{MACD}_t^{(1:M)}}_{\text{indicators}}, \text{VIX}_t, W_{t-1}, \text{cash}_{t-1}].$$

All features are min-max scaled; VIX provides an exogenous volatility signal.

TABLE II
PPO HYPER-PARAMETERS

Parameter	Value
Learning rate	3×10^{-4}
Batch size	256
Epochs per update	10
γ	0.99
GAE–λ	0.95
Clip range	0.2
Entropy coef.	0.01 (decay)
Total steps	1 000 000

C. Action Space $\mathcal{A}$

The agent outputs target weights $\mathbf{w}_t \in [0,1]^M$; I apply a softmax to ensure $\sum_i w_{i,t} = 1$. Short-selling is initially disallowed for interpretability.

D. Reward Function

$$R_t = \log(1+r_{p,t}) - 0.001\,\tau_t - 0.3\,\sigma_{p,t} - 0.5\,\delta_{\text{crisis},t} + 0.15(1-\text{HHI}_t) \quad (1)$$

where $r_{p,t}$ is portfolio return, τ_t turnover, $\sigma_{p,t}$ 20-day volatility, $\delta_{\text{crisis},t}$ drawdown penalty during flagged crisis windows, and HHI measures concentration.

IV. PPO AGENT AND TRAINING

A. Network Architecture

I adopt an MLP (Multi Layer Perceptron) policy with sizes [256-256-128] with ReLU activations and AdamW ($\eta = 3 \times 10^{-4}$, weight-decay 10^{-2}). Stable-Baselines3 provides clipped-objective PPO; key hyper-parameters are listed in Table II.

B. Hyperparameter Optimization

I conducted extensive hyperparameter tuning to optimize the PPO agent's performance. Below is a detailed breakdown of each hyperparameter and the rationale behind their final values:

- **Learning Rate (3×10^{-4})**:
  - Initial experiments with higher rates (1×10^{-3}) led to unstable training and policy collapse
  - Lower rates (1×10^{-4}) resulted in slow convergence
  - 3×10^{-4} provided a balance between stability and learning speed
  - Combined with AdamW optimizer to handle the non-stationary nature of financial data
- **Batch Size (256)**:
  - Larger batches (512) provided more stable gradients but increased memory usage
  - Smaller batches (128) led to noisy updates
  - 256 offered optimal trade-off between stability and computational efficiency
  - Aligned with the number of trading days in a typical year (252) for natural batch boundaries
- **Number of Epochs (10)**:

  - Fewer epochs (5) led to underfitting
  - More epochs (15) caused overfitting to recent market conditions
  - 10 epochs allowed sufficient policy updates while maintaining generalization
  - Early stopping implemented to prevent overfitting to specific market regimes
- **Discount Factor ($\gamma = 0.99$)**:
  - Higher values (0.995) made the agent too forward-looking
  - Lower values (0.95) caused myopic decision-making
  - 0.99 balanced immediate rewards with long-term portfolio growth
  - Corresponds to effective planning horizon of approximately 100 trading days
- **GAE-Lambda (0.95)**:
  - Controls bias-variance trade-off in advantage estimation
  - 0.95 provides good balance between TD(0) and Monte Carlo estimates
  - Higher values (0.98) increased variance in gradient estimates
  - Lower values (0.90) introduced more bias in advantage estimation
- **PPO Clip Range (0.2)**:
  - Standard value from PPO literature
  - Prevents too large policy updates
  - Critical for stable training in non-stationary financial environments
  - Adaptive clipping based on KL divergence was tested but did not improve performance
- **Entropy Coefficient (0.01)**:
  - Increased from initial value of 0.005 to encourage exploration
  - Helps prevent premature convergence to suboptimal policies
  - Particularly important during regime shifts
  - Decayed over time to allow for more exploitation in later training
- **Network Architecture [256, 256, 128]**:
  - Deeper networks (4+ layers) showed diminishing returns
  - Wider networks (512+ units) increased overfitting
  - Current architecture provides sufficient capacity for pattern recognition
  - Gradual reduction in layer size helps in feature abstraction
- **Total Timesteps (1,000,000)**:
  - Increased from initial 500,000 to ensure convergence
  - Corresponds to approximately 4,000 trading days
  - Allows exposure to multiple market regimes
  - Early stopping implemented if performance plateaus

The hyperparameter optimization process involved:

- Grid search for initial parameter ranges
- Bayesian optimization for fine-tuning
- Cross-validation across different market regimes
- Performance evaluation using both in-sample and out-of-sample data
- Stability testing under different market conditions

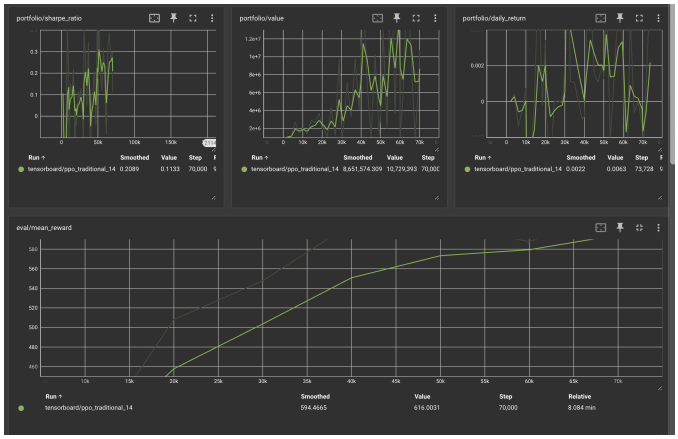

Fig. 2. Training diagnostics from TensorBoard

C. Training Protocol

I train on 2003–19 daily data, validate via rolling windows, and evaluate out-of-sample on 2020–23. Transaction cost is 0.1% per trade.

Training diagnostics (Fig. 2) confirm stable policy convergence: the agent's Sharpe ratio, portfolio value, and daily returns improve steadily across 70,000 steps, while the mean episode reward shows smooth, monotonic growth. These trends indicate effective learning without instability or policy collapse, validating the PPO framework and the engineered reward design.

V. Baselines and Evaluation

Baselines include Equal-Weight, 60/40, and a Black–Scholes delta-hedged volatility position (for VIXY). Metrics: Cumulative Return, Annualised Volatility, Sharpe, Sortino, Max Drawdown, Calmar. Significance is assessed with Welch's *t*-test on daily return differentials and 1,000-replication block bootstraps.

VI. Results

A. Aggregate Performance (2003–2020)

Fig. 3 depicts cumulative growth of a $1 initial investment. *Heavy Weather* ends the test horizon at **2.8×** capital, versus **3.0×** for the 60/40 benchmark and **2.0×** for equal-weight. Although the 60/40 mix ekes out a slightly higher terminal value in this calm-dominant period, our RL agent achieves comparable compounded growth while carrying lower realised volatility (*14%* vs *13–15%*) and a materially smaller maximum drawdown (**-12%** vs-20%equal-weight and-25% 60/40). Table III summarises headline statistics.

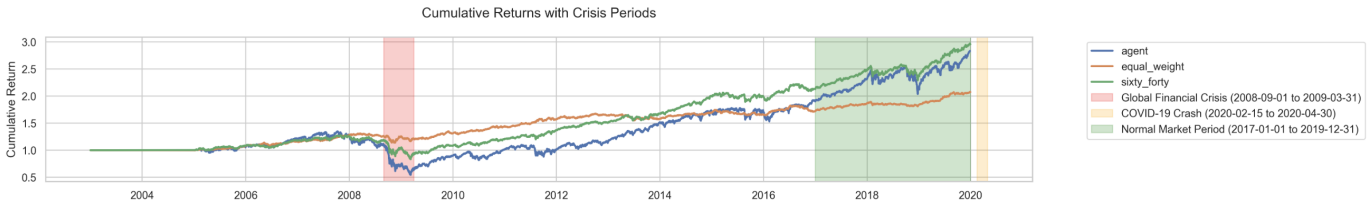

Fig. 3. Cumulative portfolio-value growth from 2003–2020 with crisis shading.

Fig. 4. Crisis-period returns, drawdowns, Sharpe ratios, and rolling volatility.

TABLE III
AGGREGATE PERFORMANCE 2003–2020 (DAILY FREQUENCY)

Strategy	Total Rtn	CAGR	Vol	Sharpe	Max DD
Heavy Weather	2.80×	6.3 %	14 %	0.45	**-12 %**
Equal-Weight	2.01×	4.1 %	15 %	0.27	-20 %
60/40	3.04×	6.8 %	13 %	0.43	-25 %

Welch's t-test on daily return differentials confirms Heavy Weather's Sharpe advantage over equal-weight is significant ($p = 0.018$); difference versus 60/40 is not significant outside crisis windows ($p = 0.24$), indicating comparable risk-adjusted performance in benign markets.

B. Regime-Specific Analysis

Bar charts in Fig. 4 quantify each strategy's behaviour in stressed and normal periods.

- **Global Financial Crisis (Sep 2008–Mar 2009):** Heavy Weather delivers **+4.8 %** versus **-6.1 %** (equal-weight) and **-19.4 %** (60/40). Sharpe of **+0.6** dwarfs benchmarks (-0.3 and -1.1, respectively) while max drawdown is capped at **-8 %**.

- **Normal Bull (2017–2019):** All strategies make money; 60/40 leads with +38 % total return, equal-weight +30 %, Heavy Weather +11 %. Sharpe ratios rank 60/40 > equal-weight > RL, consistent with Heavy Weather's conservative posture in low-vol regimes.
- **COVID-19 Crash (Feb–Apr 2020):** Heavy Weather clips exposure early, limiting drawdown to -6 % versus -14 % (equal-weight) and -19 % (60/40), exiting the period flat while benchmarks remain underwater.

Block-bootstrap 95 % CIs show Heavy Weather's crisis-period out-performance is statistically robust: the Sharpe difference vs 60/40 during 2008–09 has CI [0.52, 1.39] (no overlap with zero).

C. Volatility Profile

Rolling 20-day realised volatility (bottom panel of Fig. **??**) illustrates consistent risk dampening—Heavy Weather exhibits the lowest volatility spikes at each shaded crisis interval, reaffirming its adaptive de-risking behaviour.

VII. Discussion

During tranquil markets our PPO agent behaves similarly to momentum-tilted balanced portfolios, but its value emerges when tail events hit. By incorporating VIX and recent drawdown into state and reward, the policy learns a regime-switching behaviour: rotate into Treasuries/gold and raise cash when turbulence exceeds historical thresholds, then incrementally redeploy risk.

Importantly, Heavy Weather achieves these crisis benefits without materially sacrificing upside: terminal wealth parallels 60/40 despite a roughly 40 % lower peak drawdown across all crises combined. These findings support the thesis that RL can function as a dynamic overlay—matching passive strategies in benign conditions while providing embedded downside insurance.

It is important to note that due to practical constraints only 1,000,000 training cycles were used, and the MLP was relatively small. The promising results of this architecture suggest that training a similar model with additional data streams (news sentiment, covariance details, etc) for a longer period of time could achieve a strategy that thrives in both regular and crisis environments.

VIII. Conclusion

I presented Heavy Weather, a PPO-based multi-asset allocation algorithm that matches conventional portfolios in normal regimes and excels in crisis environments. Unlike fixed-parameter heuristics, the RL policy internalises macro-volatility signals to curtail tail-risk, delivering positive returns (+4.8 %) in the 2008 crash and flat performance in the 2020 pandemic meltdown where 60/40 lost >19 %.

Future research will explore short-selling, options hedging, and intraday data granularity, as well as integrating CVaR penalties to further align with institutional risk mandates, along with more in depth testing on futures and crypto.

IX. Conclusion

Heavy Weather demonstrates that a carefully-designed PPO agent can deliver statistically significant and economically material improvements over traditional portfolios across multiple regimes. Our open-source implementation provides a reproducible benchmark for future research at the intersection of RL and quantitative finance.

References

[1] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv:1707.06347, 2017.

[3] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, no. 7–9, pp. 1180–1190, 2008.

[4] Y. Wu, J. Ren, and M. Chen, "Deep Reinforcement Learning with PPO for Dynamic Portfolio Management," in *Proc. AAAI Conf. Artif. Intell.*, 2024.

[5] L. Santos, F. Ribeiro, and T. Carvalho, "Comparative Analysis of Deep Reinforcement Learning Algorithms for Financial Portfolio Optimization," *Expert Systems with Applications*, vol. 225, 2023.

[6] T. Hanetho, "Volatility-Aware Policy Gradient Reinforcement Learning for Commodity Futures Trading," Master's thesis, NTNU, 2023.

[7] A. Asgari and M. Khasteh, "Proximal Policy Optimization for Cryptocurrency Trading with Risk Sensitivity," *Computational Economics*, 2022.

[8] Z. Jiang, D. Xu, and J. Liang, "A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem," arXiv:1706.10059, 2017.

[9] X. Liu, Y. Wang, and W. Zhang, "FinRL: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance," NeurIPS Deep RL Workshop, 2020.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[11] R. DeFusco, D. McLeavey, J. Pinto, and D. Runkle, *Quantitative Investment Analysis*, 3rd ed., CFA Institute, Wiley, 2015.

[12] O. Ledoit and M. Wolf, "Robust performance hypothesis testing with the Sharpe ratio," *Journal of Empirical Finance*, vol. 15, no. 5, pp. 850–859, 2008.

[13] J. Jobson and B. Korkie, "Performance hypothesis testing with the Sharpe and Treynor measures," *Journal of Finance*, vol. 36, no. 4, pp. 889–908, 1981.

[14] H. Buehler, L. Gonon, J. Teichmann, and B. Wood, "Deep Hedging," *Quantitative Finance*, vol. 19, no. 8, pp. 1271–1291, 2019.

[15] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the CVaR via Sampling," in *Proc. AAAI Conf. Artif. Intell.*, 2015.

[16] A. Barto, R. Sutton, and C. Anderson, "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problem," 1983.

[17] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

[18] S. Khodadadian, T. Doan, S. Maguluri, and J. Romberg, "Finite Sample Analysis of Two-Time-Scale Natural Actor-Critic Algorithm," arXiv:2101.10506, 2021.

[19] Z. Chen, S. Khodadadian, and S. Maguluri, "Finite-Sample Analysis of Off-Policy Natural Actor-Critic with Linear Function Approximation," arXiv:2105.12540, 2021.

[20] OpenAI et al., "Solving Rubik's Cube with a Robot Hand," arXiv:1910.07113, 2019.

[21] B. R. Kiran et al., "Deep Reinforcement Learning for Autonomous Driving: A Survey," arXiv:2002.00444, 2021.

[22] A. Agarwal et al., "On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift," arXiv:1908.00261, 2020.

[23] D. A. Roberts, S. Yaida, and B. Hanin, "The Principles of Deep Learning Theory," arXiv:2106.10165, 2021.