Adaptive Reinforcement Learning Ensemble for Dynamic Portfolio Allocation

Arya Farivar
Queen's University
arya.farivar@queensu.ca

Lukas Cupsa
Queen's University
22hsq2@queensu.ca

Alex Levesque
Queen's University
alex.levesque@queensu.ca

Jacob Power
Queen's University
24vlh3@queensu.ca

Yumna Sultan
Queen's University
18yfs@queensu.ca

Shaun Thomas
Queen's University
22fff2@queensu.ca

***Abstract*—Reinforcement learning offers a promising framework for dynamic portfolio allocation, yet individual agents are often brittle across changing market regimes. This work presents an adaptive ensemble strategy that trains three deep reinforcement learning agents: Proximal Policy Optimization, Advantage Actor-Critic, and Twin Delayed Deep Deterministic Policy Gradient, on a diversified universe of nine exchange-traded funds spanning U.S. and international equities, fixed income, commodities, and real estate. A rolling 20-day Sharpe ratio selector dynamically assigns portfolio control to the best-performing agent at monthly rebalance points. Over an out-of-sample test period from October 2022 to December 2025, the ensemble achieved an annualized Sharpe ratio of 1.14 with a maximum drawdown of -10.9%, roughly half that of a passive S&P 500 benchmark. Bootstrap confidence intervals confirm that the ensemble's risk-adjusted performance is marginally statistically significant.**

I. Introduction

A. Motivation

Traditional portfolio management, such as Markowitz's Mean-Variance Optimization, often relies on static assumptions and historical covariance matrices that fail to account for the non-stationary nature of financial markets. While these frameworks struggle to adapt to regime shifts, Reinforcement Learning (RL) provides a robust alternative by framing portfolio allocation as a continuous decision-making process. RL agents learn to maximize cumulative rewards through environmental interaction, making them better suited for dynamic market conditions. Recent frameworks like FinRL [1] demonstrate that deep RL effectively navigates high-dimensional state spaces, providing a scalable solution for multi-asset allocation.

This work proposes an adaptive ensemble of three deep RL agents for daily portfolio allocation across a diversified universe of nine exchange-traded funds (ETFs).

B. Related Works

The application of Deep Reinforcement Learning (DRL) to portfolio management has advanced considerably in recent years. Actor-Critic methods such as Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C) have been shown to capture non-linear market patterns in multi-asset settings [2]. However, individual RL agents are often brittle, exhibiting high variance across market regimes. To address this, ensemble strategies that combine multiple agents have been proposed, improving robustness by allowing a system to adapt its behavior as conditions change [3].

One promising direction is adaptive model selection, where a rolling validation window is used to switch between agents based on risk-adjusted performance. Yang et al. [2] demonstrated that selecting the agent with the highest recent Sharpe ratio can outperform static benchmarks such as minimum-variance portfolios. Despite these advances, many studies focus on single-asset trading or lack rigorous statistical validation against broad market benchmarks [4]. Furthermore, open-source implementations frequently overlook realistic market frictions such as transaction costs and turnover penalties.

C. Problem Definition

The objective is to learn an optimal policy π that maps market observations to portfolio weights for a universe of $A = 9$ ETFs. At each time step t, the agent observes a matrix of trailing daily log returns with shape (W, A), where $W = 20$ is the lookback window. Each entry represents the log return of asset i at lag τ:

$$r_{t-\tau,\,i} = \log(P_{t-\tau,\,i} \,/\, P_{t-\tau-1,\,i}). \tag{1}$$

The agent outputs a real-valued vector $\mathbf{z} \in \mathbb{R}^A$, which is mapped to portfolio weights via a softmax transformation to ensure non-negative weights that sum to one:

$$\omega_{i,t} = \frac{e^{z_i}}{\sum_{j=1}^{A} e^{z_j}}. \tag{2}$$

The reward at each step is the net portfolio return after a half-spread transaction cost proportional to turnover (10 basis points):

$$R_t = \mathbf{w}_t^\top \mathbf{r}_t - c_t. \tag{3}$$

II. Methodology

A. Data and Preprocessing

We study daily portfolio allocation over a diversified universe of nine ETFs: SPY (U.S. equities), VEU (international equities),

TLT (long-duration Treasuries), TIP (inflation-protected securities), LQD (investment-grade credit), DBC (commodities), VNQ (real estate), BIL (short-term bills), and SH (inverse equity hedge). Historical adjusted close prices are obtained from Yahoo Finance over January 2010 to December 2025. We retain only the intersection of trading dates for which all assets have available data and compute daily log returns per (1).

The data are split chronologically with no overlap: training on indices $[0,\ 2413)$, validation on $[2413,\ 3217)$, and testing on $[3217,\ \text{end})$. The test period spans October 2022 to December 2025, yielding 785 out-of-sample daily returns after the 20-day lookback warm-up.

B. Trading Environment

We implement a custom Gymnasium environment for sequential portfolio allocation. At each step, the agent observes the trailing (20×9) return matrix and outputs logits mapped to portfolio weights via softmax (2), enforcing a fully invested, long-only portfolio. The reward is the net portfolio return (3) with a half-spread transaction cost of 10 basis points applied to the L_1 turnover. The environment records episode-level diagnostics including Compound Annual Growth Rate (CAGR), Sharpe ratio, maximum drawdown, and average turnover.

C. RL Agents

Three independent agents serve as base learners, all implemented using Stable-Baselines3 [5] with identical two-hidden-layer multilayer perceptron (MLP) policies (256×256 units), trained for 300,000 timesteps at a learning rate of 10^{-4}. PPO and A2C save the best checkpoint according to validation Sharpe ratio, evaluated every 10,000 training steps. TD3 uses periodic evaluation and selects the checkpoint with the highest mean validation reward. Agent-specific hyperparameters are summarized in Table I.

TABLE I
AGENT-SPECIFIC HYPERPARAMETERS. ALL AGENTS SHARE A $[256, 256]$ MLP ARCHITECTURE, 3×10^5 TIMESTEPS, LR $= 10^{-4}$, AND $\gamma = 0.99$.

Parameter	PPO	A2C	TD3
n_{steps}	2048	32	–
Batch size	128	–	256
Entropy coeff.	0.005	0.005	–
λ_{GAE}	0.95	0.95	–
Clip range	0.2	–	–
Replay buffer	–	–	2×10^5
Action noise σ	–	–	0.1

PPO [6] uses clipped surrogate objectives with 10 epochs per rollout update. A2C [7] uses short rollout buffers for fast gradient updates with gradient clipping at 0.5. Twin Delayed Deep Deterministic Policy Gradient (TD3) [8] is an off-policy actor-critic method that uses a replay buffer, delayed policy updates (every 2 steps), and Gaussian action noise for exploration.

D. Baselines

We compare against three baselines evaluated on the same test dates: an equal-weight portfolio rebalanced daily through the same trading environment and cost model, a passive buy-and-hold SPY benchmark, and a minimum-variance portfolio computed via long-only quadratic optimization on the pre-test covariance matrix.

E. Adaptive Ensemble Selection

Our core contribution is an adaptive ensemble that selects among the three RL agents at monthly rebalance points using recent risk-adjusted performance. For each agent $k \in \{\text{PPO, A2C, TD3}\}$, we compute a 20-day rolling annualized Sharpe ratio:

$$S_t^{(k)} = \sqrt{252}\ \frac{\mu_{t,20}^{(k)}}{\sigma_{t,20}^{(k)}}, \tag{4}$$

where $\mu_{t,20}^{(k)}$ and $\sigma_{t,20}^{(k)}$ are the rolling mean and standard deviation of agent k's daily returns over the prior 20 days. At each rebalance date, the ensemble selects the agent with the highest previous-day rolling Sharpe:

$$k^*(t) = \arg\max_k\ S_{t-1}^{(k)}. \tag{5}$$

Using $S_{t-1}^{(k)}$ rather than same-day values avoids look-ahead bias. Between rebalance dates, the previously selected agent continues. During the initial warm-up period, before sufficient history is available, the ensemble defaults to an equal-weight average of all three agent returns.

F. Statistical Validation

To quantify uncertainty in Sharpe ratio estimates, we compute nonparametric bootstrap confidence intervals using 10,000 resamples of the daily test returns, drawn with replacement. For each resample, the annualized Sharpe ratio is recomputed, and 95% percentile intervals are reported.

III. RESULTS

A. Overall Performance

Table II summarizes the test-period performance of all strategies. No single RL agent outperformed the passive SPY benchmark on raw returns (19.7% CAGR), though SPY carried the deepest drawdown at -19.2%. Among the individual agents, TD3 achieved the highest CAGR (10.7%) but also the highest volatility (14.6%). A2C delivered the best risk-adjusted performance of the three with a Sharpe ratio of 0.82, while PPO learned a conservative, low-volatility policy (7.5% vol) that produced modest returns.

TABLE II
TEST-PERIOD PERFORMANCE (OCT 2022 – DEC 2025).

Strategy	CAGR	Sharpe	Vol	Max DD
Ensemble	**12.5%**	**1.14**	**10.9%**	**−10.9%**
PPO	3.5%	0.49	7.5%	−9.9%
A2C	5.6%	0.82	7.0%	−8.0%
TD3	10.7%	0.77	14.6%	−17.2%
Equal-Wt	4.8%	0.84	5.7%	−6.4%
SPY B&H	19.7%	1.24	15.4%	−19.2%

A minimum-variance baseline was also evaluated but is excluded from Table II as it converged to a near-cash allocation

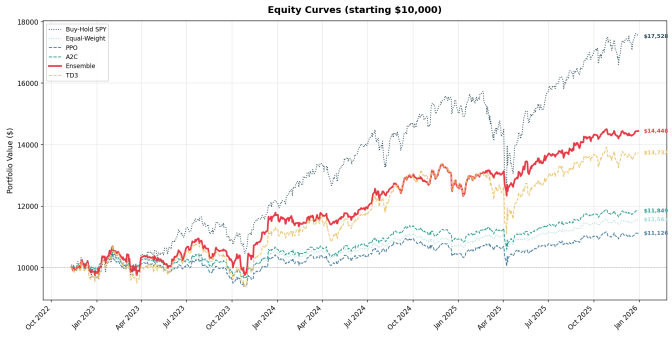

Fig. 1. Cumulative wealth curves over the test period. The ensemble (black) delivers smoother growth with lower drawdowns than SPY or any individual RL agent.

Fig. 2. Ensemble agent selection over time. TD3 is selected for the majority of the test period, with A2C chosen during periods where TD3's recent performance deteriorates.

(0.2% volatility), producing an artificially inflated Sharpe ratio of 18.17.

B. Ensemble Performance

The adaptive ensemble achieved the highest Sharpe ratio of any RL-based strategy at 1.14, exceeding every individual agent and the equal-weight baseline. Its maximum drawdown of -10.9% was roughly half that of SPY, demonstrating meaningful downside protection. Fig. 1 shows the cumulative wealth curves over the test period; the ensemble tracks upward with notably smoother growth than TD3 or SPY.

Over the test period, the ensemble selected TD3 for 53% of days, A2C for 34%, PPO for 8%, and defaulted to equal-weight averaging for the remaining 4% during the warm-up window (Fig. 2). This selection pattern reflects TD3's stronger absolute returns and confirms that the rolling Sharpe mechanism meaningfully differentiates among agents rather than cycling randomly.

Although the ensemble did not match SPY's raw CAGR (12.5% vs. 19.7%), it delivered this return at significantly lower risk. For a risk-aware allocator, the ensemble's Sharpe-to-drawdown profile represents a more favorable trade-off than passive equity exposure.

C. Statistical Significance

Table III reports 95% bootstrap confidence intervals for the annualized Sharpe ratio of each strategy. The ensemble's interval $[0.04,\ 2.26]$ is the only RL-based result whose lower bound excludes zero, providing marginal statistical evidence that its risk-adjusted outperformance is non-trivial. The individual agents: PPO, A2C, and TD3, all have confidence intervals that include zero, meaning their standalone Sharpe ratios cannot be distinguished from noise at the 95% level. The wide intervals are characteristic of approximately 785 daily observations and reflect the inherent difficulty of establishing statistical significance in financial return data.

TABLE III
BOOTSTRAP 95% CONFIDENCE INTERVALS FOR ANNUALIZED SHARPE RATIOS (10,000 RESAMPLES).

Strategy	Sharpe	95% CI
Ensemble	**1.14**	**[0.04, 2.26]**
SPY B&H	1.24	[0.17, 2.32]
Equal-Wt	0.84	[−0.26, 1.96]
A2C	0.82	[−0.28, 1.93]
TD3	0.77	[−0.32, 1.88]
PPO	0.49	[−0.60, 1.60]

D. Agent Diversity

Correlation analysis of daily returns revealed that PPO and A2C learned highly similar strategies ($\rho > 0.9$), while TD3 exhibited substantially lower correlation with both. This limited diversity partly explains the ensemble's heavy reliance on TD3 and suggests that the benefit of ensembling was constrained by redundancy between two of the three agents. Introducing greater architectural or reward-function diversity across agents is a promising direction for improving ensemble effectiveness.

IV. Conclusion

This work presented an adaptive ensemble of three deep RL agents: PPO, A2C, and TD3, for daily portfolio allocation across nine ETFs spanning distinct macro asset classes. The ensemble dynamically selected the best-performing agent at monthly rebalance points using a 20-day rolling Sharpe ratio with a previous-day lookback to prevent look-ahead bias.

The ensemble achieved a Sharpe ratio of 1.14 with a maximum drawdown of -10.9%, roughly half that of a passive SPY benchmark (-19.2%). Bootstrap analysis confirmed marginal statistical significance, with the ensemble being the only RL-based strategy whose 95% confidence interval excluded zero. However, the strategy did not surpass SPY in raw compound returns (12.5% vs. 19.7%), and correlation analysis revealed that PPO and A2C learned highly similar policies ($\rho > 0.9$), limiting the diversity available to the ensemble.

Several directions remain for future work. Training and evaluating across multiple random seeds would better characterize performance stability. The hard selection mechanism could be replaced with soft weighting, blending agent outputs proportionally to their rolling Sharpe scores. Incorporating market turbulence indicators into the rebalancing logic may further improve adaptability during volatile periods. Finally,

enriching the observation space with technical indicators or macroeconomic features could give agents richer signals for allocation decisions.

V. Acknowledgments

This work was conducted through QMIND, Queen's University's undergraduate AI organization. The authors thank QMIND for providing the project framework and collaborative environment that supported this research.

References

[1] X.-Y. Liu, H. Yang, Q. Chen, R. Zhang, L. Yang, B. Xiao, and C. D. Wang, "FinRL: A deep reinforcement learning library for automated stock trading in finance," *arXiv preprint arXiv:2011.09607*, 2020.

[2] H. Yang, X.-Y. Liu, S. Zhong, and A. Walid, "Deep reinforcement learning for automated stock trading: An ensemble strategy," in *Proceedings of the First ACM International Conference on AI in Finance (ICAIF)*, 2020.

[3] H. Zheng *et al.*, "Stock portfolio management by using fuzzy ensemble deep reinforcement learning," *Journal of Risk and Financial Management*, vol. 16, no. 3, 2023.

[4] B. Hambly, R. Xu, and H. Yang, "Recent advances in reinforcement learning in finance," *Mathematical Finance*, vol. 33, no. 3, pp. 437–503, 2023.

[5] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[7] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2016.

[8] S. Fujimoto, H. v. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning (ICML)*, 2018.