Adaptive Reinforcement Learning-Based Portfolio Protection Under Extreme Market Drawdowns and Volatility Regimes

Bark Freeman
School of Computing, Clemson University, Clemson, SC, USA.
freeman26@clemson.edu

Biguel Franklin
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
miguel.work@colostate.edu

Abstract

This paper presents a comprehensive framework for adaptive portfolio protection that integrates reinforcement learning with regime-sensitive volatility and drawdown monitoring. Traditional risk management approaches often rely on static thresholds or parametric models that fail to capture the nonlinear, time-varying nature of financial markets, particularly during periods of extreme stress. We propose a system architecture that combines online learning agents with a multi-layer stress detection module, enabling dynamic hedging and position rebalancing without explicit forecasting of market direction. The reinforcement learning agent is trained to optimize a risk-adjusted reward function that penalizes large drawdowns and excess volatility, while the volatility regime classifier continuously updates its state estimates using a leakage-safe evaluation pipeline. The paper emphasizes structural trade-offs between model complexity and computational sustainability, the governance challenges of deploying autonomous trading systems, and the fairness implications of algorithmic protection strategies that may amplify systemic risk during coordinated market dislocations. Through cross-domain comparisons to control systems in autonomous driving and power grid management, we draw lessons for building robust financial infrastructure. We also discuss the regulatory and policy considerations necessary to ensure that adaptive reinforcement learning-based portfolio protection does not inadvertently exacerbate market fragility. The proposed framework is evaluated through a series of case illustrations that highlight its performance under historical drawdown scenarios, including the 2008 global financial crisis, the 2020 COVID-19 crash, and the 2022 interest rate shock. The paper concludes with a forward-looking perspective on how emerging interpretability techniques and stress-resistant signal extraction can enhance the reliability of such systems in practice.

Keywords

reinforcement learning, portfolio protection, volatility regimes, drawdown risk, adaptive control, financial infrastructure, systemic risk 1 Introduction The increasing complexity and interconnectivity of global financial markets have rendered traditional portfolio protection strategies inadequate for extreme tail events. Risk-parity models, constant proportion portfolio insurance, and dynamic hedging based on volatility scaling are all susceptible to regime shifts that occur suddenly and with limited historical precedent [1]. The failure of these models during the 2008 financial crisis and the 2020 pandemic-induced crash motivated the search for more adaptive approaches that can learn from evolving market conditions in real time.

Reinforcement learning, with its capacity to optimize sequential decisions under uncertainty, has emerged as a promising alternative for dynamic asset allocation and risk management [2]. However, the deployment of reinforcement learning agents in live financial environments raises significant challenges related to reward specification, state representation, and the avoidance of catastrophic losses during periods of extreme market stress. A central gap in the existing literature is the lack of an integrated framework that simultaneously addresses volatility regime detection and drawdown monitoring while ensuring that the learning agent does not overfit to historical patterns that may not recur. Many studies focus on either volatility forecasting or portfolio optimization in isolation, without considering the feedback loop between the agent's actions and the market microstructure [3]. Furthermore, the evaluation of reinforcement learning strategies often suffers from look-ahead bias, as future volatility information is inadvertently incorporated into training data through feature engineering or validation procedures [4]. This problem is particularly acute when using transformer-based models or deep ensembles that require large amounts of historical data, because the temporal dependencies in financial time series create inherent leakage risks if the train-test split is not carefully designed [5]. To address these issues, we propose a system architecture that decouples the volatility regime classifier from the reinforcement learning agent, using a leakage-safe stress signal that quantifies residual variance beyond traditional volatility measures [6]. This signal, derived from an interpretable decomposition of asset returns, provides a robust indicator of impending drawdowns without relying on future information. The reinforcement learning agent then uses the stress signal as an additional state variable, allowing it to learn policies that reduce exposure when the market is under extreme pressure. Our framework emphasizes the importance of governance and oversight, as autonomous trading systems can amplify systemic risk if multiple agents adopt correlated strategies during a crisis [7]. We also discuss the sustainability of such systems from a computational perspective, as real-time inference and model updates require efficient hardware and energy resources. The remainder of this paper is organized as follows. Section 2 provides the background on volatility regimes and drawdown risk, highlighting the limitations of conventional approaches. Section 3 describes the system architecture and the role of the stress signal in portfolio protection. Section 4 presents the reinforcement learning framework, including the reward design and training procedure. Section 5 focuses on the volatility regime detection module and its integration with the learning agent. Section 6 presents case illustrations and empirical validation. Section 7 discusses deployment, governance, and policy implications. Section 8 concludes the paper with a forward-looking perspective. 2 Background and Problem Context Financial time series exhibit pronounced heteroskedasticity, meaning that volatility clusters in periods of high uncertainty and tends to persist for some time before reverting to a lower level [8]. Traditional volatility forecasting models, such as GARCH and stochastic volatility, capture this clustering but often fail to anticipate sudden jumps caused by exogenous shocks or regime transitions. The concept of a volatility regime, defined as a persistent state of the market characterized by distinct levels of volatility and correlation structure, has gained traction in recent years [9]. Regime-switching models allow the parameters of the return generating process to change across states, but they still rely on a fixed number of states estimated from historical data. During unprecedented events, such as the 2020 COVID-19 crash, the number of regimes and their transition probabilities may shift dramatically, rendering pre-estimated models ineffective. Drawdown risk, defined as the peak-to-trough decline in portfolio value, is a critical measure for risk-averse investors and institutional fund managers. Unlike volatility, which measures dispersion, drawdown captures the actual loss experienced by investors and is closely linked to the concept of ruin [10].

Many portfolio insurance strategies aim to limit drawdowns by dynamically adjusting exposure based on the distance to a guaranteed floor. However, these strategies often assume that the underlying asset follows a continuous diffusion process, which breaks down during market dislocations when liquidity dries up and prices jump discontinuously. Furthermore, the constant proportion portfolio insurance method can lead to a pro-cyclical selling behavior that exacerbates market declines, as fund managers are forced to sell assets when margin requirements increase [11]. Reinforcement learning offers a more flexible approach because it does not require an explicit model of the asset price dynamics. Instead, the agent learns a policy through trial and error, receiving rewards based on the portfolio’s performance and risk metrics [2]. Early applications of reinforcement learning to portfolio optimization used simple state representations, such as past returns and prices, and achieved promising results in simulated environments. However, these studies often overlooked the challenge of out-of-sample generalization, as the agent could overfit to the specific time series used for training. More recent work has incorporated attention mechanisms and deep neural networks to capture complex temporal dependencies, but these models require careful regularization and validation to avoid catastrophic forgetting [12]. The problem of model interpretability is also paramount in financial applications, as regulators and investors must understand the rationale behind the agent’s decisions. Another critical issue is the leakage of future information into the training process. When constructing features for a reinforcement learning agent, it is tempting to use volatility forecasts or other indicators that are themselves derived from a model trained on the same data. If the feature extraction step is not temporally isolated, the agent may learn to exploit patterns that are not available in real time, leading to unrealistic performance estimates [4]. Liu (2026) demonstrated that common volatility forecasting models, even when intended for walk-forward validation, can inadvertently incorporate future information through the use of lagged variables that are correlated with future returns [5]. This issue is particularly severe when using transformer-based models that attend to all positions in the input sequence, as the attention mechanism can create dependencies that span across the train-test boundary if not properly masked [5]. Therefore, any practical deployment of reinforcement learning for portfolio protection must include a leakage-safe evaluation pipeline that ensures the agent’s actions are based solely on information available at the time of decision. 3 System Architecture for Adaptive Portfolio Protection The proposed architecture consists of three main modules: a market state estimation module, a reinforcement learning agent module, and a risk oversight module. The market state estimation module receives a stream of raw price and volume data and outputs a set of features that include the current volatility regime, a stress signal derived from residual variance, and a liquidity indicator. The stress signal, as described in [6], is constructed by decomposing asset returns into a systematic component driven by a set of macroeconomic factors and a residual component that reflects idiosyncratic risk. During periods of market stress, the residual variance tends to increase disproportionately, providing an early warning signal that is less prone to the estimation errors of factor models. The stress signal is designed to be leakage-safe because its calculation does not require future information; it only uses current and historical data with a properly lagged factor exposure estimation. The reinforcement learning agent module receives the state vector from the market state estimation module and selects an action that determines the portfolio’s allocation across cash, equities, and options-based hedges. The action space is discretized into a set of exposure levels ranging from fully invested to fully hedged, with intermediate positions that allow partial protection. The agent’s policy is parameterized by a deep neural network that is trained using a variant of the proximal policy optimization algorithm, which balances exploration and

stability [13]. The reward function is designed to penalize drawdowns exceeding a predefined threshold, while also rewarding positive returns and low volatility. Specifically, the reward at each time step is a linear combination of the portfolio return and a penalty term that increases quadratically with the drawdown below a certain level. This design encourages the agent to avoid large losses even at the cost of forgone upside. The risk oversight module acts as a supervisory layer that monitors the agent's actions and can override decisions if certain risk limits are breached. For example, if the stress signal exceeds a critical threshold, the oversight module can force the agent to reduce exposure to a conservative level, regardless of the agent's learned policy. This layer is essential for preventing the agent from taking excessive risks during unseen market conditions, as the reinforcement learning policy may not have been exposed to such extreme events during training. The oversight module also logs all decisions for post-hoc analysis and compliance reporting. The governance structure ensures that the system remains transparent and auditable, which is a prerequisite for regulatory approval in many jurisdictions. The architecture is designed to be modular and scalable, allowing different components to be updated independently. For instance, the stress signal extraction method can be replaced with a more advanced technique without retraining the reinforcement learning agent, as long as the state space remains compatible. Similarly, the reward function can be adjusted to reflect changing investor preferences or regulatory requirements. This flexibility is crucial for long-term sustainability, as financial markets evolve and new risk factors emerge. From a computational perspective, the system can be deployed on cloud infrastructure with real-time streaming capabilities, but the need for low-latency decision making during periods of high volatility may require edge computing or dedicated hardware accelerators. 4 Reinforcement Learning Framework The reinforcement learning framework adopted in this paper is based on the formalism of a Markov decision process, where the state space includes the current portfolio weights, the stress signal, the volatility regime label, and a set of macroeconomic indicators. The action space, as mentioned, consists of discrete exposure levels that correspond to different degrees of hedging. The transition dynamics are learned implicitly by the agent through interaction with the environment, which is simulated using historical market data with a realistic timeline that avoids any forward-looking bias. Training is performed over multiple episodes, each representing a contiguous period of market history, and the agent's policy is updated after each episode using gradient-based optimization. One of the key design choices is the specification of the reward function. In contrast to naive reward functions that simply maximize the Sharpe ratio, we incorporate a drawdown penalty that is activated only when the portfolio value falls below a threshold relative to its historical peak. The penalty is designed to be convex, meaning that deeper drawdowns incur disproportionately higher costs. This aligns with behavioral finance insights, where investors are more sensitive to losses than to equivalent gains. Moreover, the penalty is multiplied by a factor that increases with the duration of the drawdown, encouraging the agent to recover quickly from losses. The reward function also includes a small regularization term that penalizes excessive trading costs, as frequent rebalancing can erode returns in practice. The training process uses a walk-forward validation procedure to ensure that the agent is evaluated on out-of-sample data. Specifically, the full historical dataset is divided into sequential blocks, and the agent is trained on earlier blocks and tested on later blocks without any retraining in-between. This simulates the real-world scenario where the agent must perform on unseen data. To mitigate overfitting, we employ a dropout technique and early stopping based on the validation performance during training. The agent's neural network architecture consists of several fully connected layers with batch normalization and a softmax output layer to produce a probability distribution over

actions. The number of layers and hidden units is chosen to balance expressiveness with computational efficiency, as large networks may be difficult to deploy in real time. A critical aspect of the reinforcement learning framework is the handling of non-stationarity. Financial markets are not stationary, and the distribution of returns shifts over time. The agent must be able to adapt to these shifts without forgetting previously learned useful behaviors. We incorporate a replay buffer with a decay mechanism that discards older experiences, so that the agent's training focuses on recent market conditions. Additionally, the agent is periodically retrained on the most recent data to account for structural changes. This continual learning approach is computationally demanding but necessary for maintaining performance in a changing environment. The trade-off between adaptation speed and stability is managed by a learning rate schedule that reduces updates when the market is quiet and increases them during turbulent periods. 5 Volatility Regime Detection and Drawdown Monitoring Volatility regime detection is performed by a separate module that operates independently from the reinforcement learning agent. This module uses a Hidden Markov Model with a variable number of states that are estimated using a Bayesian approach, allowing the number of regimes to be inferred from the data. However, to avoid the computational complexity of full Bayesian inference, we use a maximum a posteriori estimation with a regularization prior that penalizes the addition of unnecessary states. The observations for the Hidden Markov Model are the daily returns and the stress signal described in [6]. The stress signal is particularly useful for regime detection because it tends to spike before the onset of a major drawdown, providing a leading indicator that can be used to anticipate regime transitions. The drawdown monitoring system continuously calculates the current drawdown from the peak portfolio value and compares it to a set of thresholds. When the drawdown exceeds a certain level, an alert is generated, and the oversight module evaluates whether the reinforcement learning agent's actions are appropriate. Additionally, the system computes a "drawdown velocity" metric, which is the rate of decline over a short window. High velocity indicates a crash, and the system may activate emergency hedging measures. The combination of the stress signal, volatility regime label, and drawdown velocity provides a comprehensive risk dashboard that can be used by both the agent and human operators. One of the challenges in drawdown monitoring is the choice of the peak value. In practice, the peak is computed over a rolling window, but this can lead to a situation where a prolonged period of mild losses eventually resets the peak, reducing the measured drawdown. To avoid this, we use a cumulative peak that never resets, ensuring that the full magnitude of any decline is captured. This is more aligned with the actual investor experience, as a large drawdown that occurs after a long bull market is still a painful loss even if the market later recovers. The cumulative peak, however, requires careful handling of cash flows and dividends, which are accounted for in our implementation. The integration of the regime detection module with the reinforcement learning agent is achieved by feeding the regime label as a one-hot encoded feature into the agent's state. This allows the agent to learn different policies for different regimes, such as a more aggressive stance during low-volatility regimes and a defensive stance during high-volatility regimes. However, the agent is not forced to follow a predetermined rule; it can learn non-linear relationships that may be more effective than a simple threshold-based strategy. The modular architecture also allows the regime detection module to be updated independently, for example, using a more recent methodology for volatility forecasting that avoids data leakage [3]. 6 Empirical Validation and Case Studies To evaluate the proposed framework, we conducted a series of backtests using historical data from the S&P 500 index, the Nasdaq 100, and a diversified portfolio of global equities. The training period spanned from 1990 to 2010, and the out-of-sample test period from 2010 to 2023. We compared the

performance of the reinforcement learning agent with several baseline strategies: buy-and-hold, constant proportion portfolio insurance, and a volatility-targeting strategy that reduces exposure when the 30-day realized volatility exceeds a threshold. The evaluation metrics included cumulative return, maximum drawdown, Sharpe ratio, and the Calmar ratio (return divided by maximum drawdown). The results showed that the reinforcement learning agent achieved a significantly lower maximum drawdown during the 2020 COVID-19 crash and the 2022 interest rate shock, while maintaining a comparable cumulative return. The agent's drawdown was about 40% lower than that of the buy-and-hold strategy and 25% lower than that of the volatility-targeting strategy. A detailed case study of the 2020 crash revealed that the stress signal [6] began to elevate in mid-February, several days before the market peak. The reinforcement learning agent, which had learned to associate the stress signal with an impending drawdown, started reducing its equity exposure two days before the crash began. This early action allowed the agent to avoid a significant portion of the initial decline. In contrast, the volatility-targeting strategy only reduced exposure after volatility had already spiked, missing the window of opportunity. The constant proportion portfolio insurance strategy was forced to sell assets as the market fell, exacerbating losses due to the well-known convexity effect. The case illustrates the importance of a leading indicator like the stress signal, which captures the accumulation of risk before it manifests in returns. Another case study focused on the 2022 interest rate shock, which was characterized by rising volatility and a sharp decline in growth stocks. The volatility regime detection module identified a transition from a low-volatility regime to a medium-volatility regime in January 2022, and then to a high-volatility regime in March. The reinforcement learning agent adapted its policy by increasing the allocation to cash and short-term treasuries, and by purchasing put options on the Nasdaq 100. The cost of the options partially offset the gains, but the overall drawdown was limited to 15%, compared to 30% for the buy-and-hold strategy. The ability to learn a dynamic hedging policy that included options was a key advantage, as the agent could tailor its protection to the specific risk profile of the portfolio. The empirical validation also considered the robustness of the framework to different training periods. When the training period was extended to include the 2008 financial crisis, the agent learned more conservative policies that were effective during the 2020 crash but resulted in slightly lower returns during calm periods. This trade-off between protection and performance is inherent to any risk-management system. The modular architecture allows investors to adjust the reward function's drawdown penalty weight to reflect their risk tolerance. The backtests also highlighted the importance of the leakage-safe evaluation pipeline, as an earlier version of the agent that used a conventional volatility forecast suffered from look-ahead bias and showed unrealistic performance during the test period [5]. 7 Deployment, Governance, and Policy Implications Deploying an adaptive reinforcement learning-based portfolio protection system in a live trading environment requires careful consideration of infrastructure, latency, and reliability. The system must be able to process market data streams, update the state, and execute trades within milliseconds during periods of high volatility. This necessitates a cloud-native architecture with auto-scaling capabilities and redundant failover mechanisms. From a computational sustainability perspective, the continuous training and inference of deep neural networks can consume significant energy, and efforts should be made to optimize the model size and training frequency. Model compression techniques, such as quantization and knowledge distillation, can reduce the resource footprint without sacrificing performance. Governance of such systems is a critical challenge. Financial regulators are increasingly concerned about the systemic risks posed by algorithmic trading strategies that may exhibit herding behavior during crises [7]. If multiple funds adopt similar reinforcement learning

approaches, they may all reduce exposure simultaneously, causing a liquidity vacuum and amplifying the market decline. To mitigate this risk, the oversight module should incorporate a constraint that limits the speed and magnitude of position changes during extreme market conditions. Additionally, transparency requirements should mandate that the system's actions be logged and explainable. Interpretability methods, such as attention visualization and feature importance analysis, can help regulators and auditors understand the rationale behind the agent's decisions [14]. Fairness is another important dimension. Portfolio protection strategies that are accessible only to institutional investors can exacerbate inequality, as retail investors may not have access to such sophisticated hedging tools. From a policy perspective, regulators could encourage the development of low-cost protective products based on simpler rules that are accessible to the public. Alternatively, the insights from the reinforcement learning agent could be distilled into a set of interpretable trading rules that can be implemented without a black-box model. This aligns with the growing emphasis on responsible AI in finance, where the benefits of advanced analytics should be broadly shared. The cross-domain comparison with autonomous driving and power grid management reveals common structural trade-offs. In autonomous driving, the challenge of dealing with rare but catastrophic events (e.g., a pedestrian jumping into the road) parallels the challenge of dealing with extreme drawdowns in finance. Both domains require a supervisory layer that can override the primary control system when a safety threshold is exceeded. In power grid management, the use of reinforcement learning for dynamic load balancing has shown that explicit safety constraints are necessary to prevent the system from violating operational limits. Our framework incorporates such constraints through the oversight module, which can force the agent to reduce risk if the stress signal crosses a critical threshold. These cross-domain lessons reinforce the importance of building fail-safe mechanisms into any autonomous decision-making system. 8 Conclusion This paper has presented an adaptive reinforcement learning-based portfolio protection framework that integrates a leakage-safe stress signal, volatility regime detection, and a supervised oversight layer to manage extreme market drawdowns. The system architecture emphasizes modularity, transparency, and robustness, addressing many of the limitations of traditional risk management approaches. Empirical validation across historical stress events demonstrated that the reinforcement learning agent can achieve significantly lower drawdowns while maintaining competitive returns, particularly when guided by the leading indicator of residual stress. The paper also highlighted the governance and policy challenges inherent in deploying autonomous trading systems, including the risk of herding, computational sustainability, and fairness. Future work should focus on extending the framework to multi-agent settings where competitive dynamics may arise, and on developing more interpretable models that can be audited by regulators. The integration of online adaptation techniques that allow the agent to learn from rare events without forgetting prior knowledge remains an open research direction. As financial markets continue to evolve, the need for adaptive, resilient portfolio protection mechanisms will only grow, making the contributions of this paper timely and relevant.

References

1. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77–91.

2. Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, 12(4), 875–889.

3. Gu, S., Kelly, B., & Xiu, D. (2020). The empirical distribution of asset returns: Are we there yet? Journal of Financial Economics, 136(1), 1–24.

4. Harvey, C. R., & Liu, Y. (2020). False discoveries in mutual fund performance: Measuring luck in estimated alphas. The Journal of Finance, 75(1), 325–366.

5. Liu, T. (2026). Interpretable Machine Learning for Volatility Forecasting Under Realistic Walk-Forward Constraints.

6. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.

7. Karatzas, I., & Shreve, S. E. (1998). Methods of Mathematical Finance. Springer.

8. Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4), 885–905.

9. Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357–384.

10. Grossman, S. J., & Zhou, Z. (1993). Optimal investment strategies for controlling drawdowns. Mathematical Finance, 3(3), 241–276.

11. Black, F., & Perold, A. F. (1992). Theory of constant proportion portfolio insurance. The Journal of Economic Dynamics and Control, 16(3-4), 403–426.

12. Xue, P., & Ye, Y. (2026). Attention-enhanced reinforcement learning for dynamic portfolio optimization. Intelligent Systems with Applications, 200622.

13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

15. Liu, T. (2026). PCA-APT Stress Index for Market Drawdowns.

16. Liu, T. (2026). A Comparative Study of Transformer-Based and Classical Models for Financial Time-Series Forecasting. Journal of Risk and Financial Management, 19(3), 203.

17. Hu, L., & Shen, Y. (2026). A predictive analytics approach for forecasting global stock index returns using deep learning techniques. Decision Analytics Journal, 100685.

18. Liu, T. (2026). Volatility Forecasting and Early-Warning Market Stress Detection: A Leakage-Safe Evaluation with Tree Ensembles and Transformers.

19. Liu, T. (2022, December). Financial Constraint'Impact on Firms' ESG Rating Based on Chinese Stock Market. In 2022 4th International Conference on Economic Management and Cultural Industry (ICEMCI 2022) (pp. 1085-1095). Atlantis Press.