# Risk-Aware Deep Reinforcement Learning with Hierarchical Adaptation for XAU/USD Trading

**Abstract**  
 XAU/USD trading is the process of speculating on the future price movements of gold (XAU) against the US dollar (USD). This speculation relies on historical price data, which exhibits constantly evolving dynamics due to shifting market conditions. Traditional trading, limited by human biases, struggles with inconsistent decisions, poor risk management, and adapting to volatile markets, while on-policy RL methods like PPO and A3C fail to leverage historical data for generalizable patterns, critical for adapting to shifting conditions with limited data. And discrete-action algorithms cannot dynamically adjust trade and leverage sizes, and existing methods lack robust exploration to handle changing market regimes (trending, volatile, flat)

To address these limitations, we propose a hierarchical reinforcement learning framework using Soft Actor-Critic (SAC) to enable adaptive, risk-conscious trading. We begin by setting up the trading environment with 1-hour XAU/USD candlestick data (2010-2020), incorporating transaction costs and slippage for realism. State representations are constructed using Gramian Angular Field heatmaps to capture price patterns, volatility indicators like ATR to quantify market conditions, and binary flags for economic events to signal regime shifts, all normalized for stability. The low-level SAC agent is trained on these states to optimize trading actions (position size, leverage) using a risk-aware reward function balancing profit, drawdown, and exploration, while the meta-controller learns on meta-states, such as aggregated volatility or CUSUM test results, to detect market regime shifts (trending, volatile, flat) and select sub-policies (e.g., adjusting leverage for aggressive or conservative mode) to guide the SAC agent. Hyperparameters are optimized via Optuna to maximize Sharpe ratio.

Subsequently, online adaptation is enabled through periodic or triggered retraining on 2020-2025 data, updating the replay buffer to incorporate recent market dynamics for continuous learning. The framework is then backtested on 2020-2025 data across trending, volatile, and flat regimes, evaluating Sharpe ratio and maximum drawdown, and compared against benchmarks like moving average crossovers (rule-based), DQN (discrete-action), and PPO (on-policy) to ensure robust performance in non-stationary markets.

This work offers a scalable framework for volatile assets like XAU/USD, with potential for real-time trading and extension to other markets, advancing algorithmic trading research and practice.

## 1. Introduction

XAU/USD trading, the speculation on gold prices against the US dollar, is a critical activity in global financial markets, driven by gold’s role as a safe-haven asset amid economic uncertainty, geopolitical tensions, and monetary policy shifts. This pair exhibits non-stationary dynamics, with prices fluctuating across trending, volatile, and flat regimes. Traditional trading approaches rely on technical analysis (e.g., moving averages, RSI) to identify price patterns and fundamental analysis (e.g., interest rates, inflation data) to assess macroeconomic drivers. However, these methods are limited by human biases, leading to inconsistent decisions, poor risk management, and challenges in adapting to rapidly shifting market conditions. In recent times, algorithmic trading has gained prominence, leveraging computational models to execute high-frequency trades with precision, yet many algorithms struggle with dynamic risk adjustment and real-time adaptation to market volatility.

Machine learning, particularly reinforcement learning (RL), has been applied across domains like robotics, gaming, and healthcare, with increasing adoption in financial markets for tasks such as portfolio optimization and trading strategy development. Despite this progress, existing RL methods face significant shortcomings in XAU/USD trading. On-policy algorithms like Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A3C) fail to efficiently reuse historical data, a critical issue given the limited and expensive nature of high-quality financial datasets (e.g., tick-level XAU/USD data). Discrete-action methods like Deep Q-Networks (DQN) cannot dynamically adjust trade sizes or leverage, essential for risk-aware trading. Moreover, these approaches lack robust exploration, often converging to deterministic policies unsuitable for non-stationary markets. Effective trading requires entropy regularization to balance exploration and reward maximization, alongside risk-aware strategies that mitigate drawdowns. Additional challenges include high transaction costs, slippage, overfitting to historical patterns, sensitivity to market noise, and the need to incorporate fundamental events (e.g., FOMC announcements). To address these limitations, we propose a hierarchical RL framework using Soft Actor-Critic (SAC), combining a low-level agent for continuous action optimization with a meta-controller for regime-adaptive policy selection, advancing risk-conscious trading in volatile markets like XAU/USD.

## 2. Related Work

The application of reinforcement learning (RL) to financial trading has garnered significant attention due to its ability to model complex, dynamic environments. Traditional trading strategies, such as moving average crossovers and trend-following rules, rely on static technical indicators but struggle with adaptability in non-stationary markets like XAU/USD [1]. Fundamental analysis, incorporating macroeconomic factors, adds context but is limited by human interpretation and latency in response to market shifts.

Early RL applications in finance used Q-learning and Deep Q-Networks (DQN) for discrete-action trading strategies [2]. While effective in structured environments, DQN’s discrete action spaces fail to capture the continuous nature of trade sizing and leverage adjustment critical for risk management. On-policy RL methods, such as Proximal Policy Optimization (PPO) [3] and Advantage Actor-Critic (A3C) [4], improve policy learning but suffer from sample inefficiency, discarding historical data after each episode, which is problematic given the high cost and scarcity of financial datasets. These methods also lack robust exploration, often converging to suboptimal policies in volatile markets.

Soft Actor-Critic (SAC) [5] addresses these issues through off-policy learning and entropy regularization, promoting exploration and continuous action spaces suitable for dynamic trading decisions. Hierarchical RL frameworks, as explored in [6], introduce meta-controllers to manage non-stationarity by selecting sub-policies for different tasks or regimes, offering a promising approach for market adaptation. Financial applications of RL, such as [7], leverage deep learning for signal representation, using features like price patterns and volatility indicators to inform trading decisions.

Despite these advances, existing methods rarely combine hierarchical structures with risk-aware rewards and continuous actions, nor do they adequately address financial-specific challenges like transaction costs, slippage, and economic event impacts. Our proposed framework builds on SAC and hierarchical RL to develop a scalable, adaptive solution for XAU/USD trading, incorporating rich state representations (e.g., Gramian Angular Field heatmaps) and regime-adaptive policies to overcome these limitations.

## 3. Methodology

This section outlines the proposed hierarchical reinforcement learning framework for XAU/USD trading, detailing the environment setup, state and action spaces, reward function, and the Soft Actor-Critic (SAC) algorithm with its hierarchical structure. We describe the trading environment, constructed from historical XAU/USD data, followed by the state representations combining technical and fundamental features for robust market modeling. The action space enables continuous trading decisions, optimized by a risk-aware reward function. The SAC algorithm is introduced, with a low-level agent handling trade execution and a meta-controller managing regime shifts. Hyperparameter optimization ensures performance, addressing the non-stationary dynamics of XAU/USD markets

### 3.2 State Representations

The state at time is a multi-modal vector capturing price dynamics, volatility, and fundamental events, designed to provide a comprehensive view of the XAU/USD market:

* **Gramian Angular Field (GAF) Heat maps**: The past 50 hourly closing prices are transformed into a 32x32 GAF image using the pyts library [1]. GAF encodes temporal correlations as angular differences, suitable for convolutional neural network (CNN) processing, capturing complex price patterns.
* **Volatility Indicators**: The 14-period Average True Range (ATR) quantifies market volatility, normalized by dividing by 100 to ensure numerical stability. Bollinger Bands (20-period, 2 standard deviations) provide additional volatility context.
* **Economic Event Flags**: Binary flags indicate high-impact events (e.g., FOMC meetings, Non-Farm Payroll releases), sourced from economic calendars, signaling potential regime shifts.

The state vector:

It is normalized using z-score to ensure consistency across features. A CNN extracts features from the GAF heatmap, concatenated with ATR and flags, yielding a high-dimensional state for the SAC agent.

### 3.3 Action Space and Continuous Actions

### The action space is continuous, defined as:

### Where:

### represents the position size and direction (negative for short, positive for long, zero for neutral).

### denotes leverage, allowing dynamic risk adjustment. Continuous actions, enabled by SAC’s Gaussian policy, allow fine-grained control over trade sizes and leverage, unlike discrete-action methods (e.g., DQN), which are limited to fixed choices. This flexibility is critical for risk-aware trading, adapting to market conditions without being constrained by predefined action bins.

### 3.4 Reward Function

### The reward function is designed to balance profitability, risk, and exploration:

### 

### Where:

### is the profit/loss from the trade at time t, calculated as:

### is the maximum drawdown, penalizing large capital loses, computed as:

### is the policy entropy, encouraging exploration to avoid overfitting to specific market patterns.

### With weights:

### That are tuned to prioritize profit while mitigating risk and promoting robustness. This risk-aware reward addresses the need for dynamic risk management in volatile XAU/USD markets.

### 3.5 Soft Actor-Critic Algorithm

### SAC is an off-policy, maximum entropy RL algorithm that optimizes a stochastic policy by balancing expected rewards and entropy [2]. The objective is:

### where is the temperature parameter controlling exploration. SAC uses two Q-networks, a policy network, and a replay buffer for sample-efficient learning. The actor (policy) outputs a Gaussian distribution over actions, while the critic estimates Q-values. This approach ensures robust exploration and continuous action optimization, ideal for non-stationary financial environments.

#### 3.5.1 Low-Level SAC Agent

### The low-level SAC agent operates within the trading environment, taking states s\_t and producing continuous actions. The policy network is a multi-layer perceptron (MLP) with a CNN front-end to process GAF heatmaps, followed by fully connected layers for ATR and event flags. The agent is trained using a replay buffer of size , storing transitions

### The Q-networks are updated via soft Q-learning, minimizing:

### The policy is updated to maximize , with alpha automatically tuned. Training occurs over timesteps, with a batch size of 256, ensuring convergence on 2010–2020 data.

#### 3.5.2 Meta-Controller

### The meta-controller, implemented as a higher-level SAC policy, addresses non-stationarity by detecting market regimes (trending, volatile, flat) and selecting sub-policies for the low-level agent. Meta-states include aggregated volatility (e.g., 100-period ATR) and CUSUM test results for structural breaks:

### The meta-controller outputs a discrete action selecting one of three sub-policies:

### Aggressive: high leverage,

### Conservative: low leverage,

### Neutral: no trade

### It is trained on meta-transitions every 100 timesteps, using a separate replay buffer. This hierarchical structure ensures adaptability to market shifts, enhancing robustness.

### 3.6. Hyperparameter Optimization

### Hyperparameters (learning rates, alpha, network sizes) are optimized using Optuna [3] to maximize the Sharpe ratio on a validation set (2019–2020 data). The optimization runs 50 trials, balancing computational efficiency and performance.

**References**

[1] Murphy, J. J. (1999). *Technical Analysis of the Financial Markets*. New York Institute of Finance.

[2] Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653-664.

[3] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

[4] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning (ICML)*, 1928-1937.

[5] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 1861-1870.

[6] Nachum, O., Gu, S., Lee, H., & Levine, S. (2018). Data-efficient hierarchical reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 3303-3313.

[7] Deng, Y., et al. (2017). As above.