Sarkany, Attila; Janásek, Lukáš; Baruník, Jozef

Working Paper

Quantile preferences in portfolio choice: A Q-DRL approach to dynamic diversification

IES Working Paper, No. 21/2024

Provided in Cooperation with:
Charles University, Institute of Economic Studies (IES)

Suggested Citation: Sarkany, Attila; Janásek, Lukáš; Baruník, Jozef (2024) : Quantile preferences in portfolio choice: A Q-DRL approach to dynamic diversification, IES Working Paper, No. 21/2024, Charles University in Prague, Institute of Economic Studies (IES), Prague

This Version is available at:
https://hdl.handle.net/10419/300178

INSTITUTE
OF ECONOMIC STUDIES
Faculty of Social Sciences
Charles University

QUANTILE PREFERENCES IN PORTFOLIO CHOICE: A Q-DRL APPROACH TO DYNAMIC DIVERSIFICATION

Attila Sarkany
Lukáš Janásek
Jozef Baruník

IES Working Paper 21/2024

Institute of Economic Studies,
Faculty of Social Sciences,
Charles University in Prague

[UK FSV – IES]

Opletalova 26
CZ-110 00, Prague
E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Institut ekonomických studií
Fakulta sociálních věd
Univerzita Karlova v Praze

Opletalova 26
110 00 Praha 1

E-mail : ies@fsv.cuni.cz
http://ies.fsv.cuni.cz

Bibliographic information:
Sarkany A., Janásek L., Baruník J. (2024): " Quantile Preferences in Portfolio Choice: A Q-DRL Approach to Dynamic Diversification" IES Working Papers 21/2024. IES FSV. Charles University.

Quantile Preferences in Portfolio Choice: A Q-DRL Approach to Dynamic Diversification

Attila Sarkany[1,2]
Lukáš Janásek[1,2]
Jozef Baruník[1,2]

[1]Institute of Economic Studies, Charles University, Prague, Czech Republic
[2]The Czech Academy of Sciences, IITA, Prague, Czech Republic E-mail:

May 2024

Abstract:
We develop a novel approach to understand the dynamic diversification of decision makers with quantile preferences. Due to unavailability of analytical solutions to such complex problems, we suggest to approximate the behavior of agents with a Quantile Deep Reinforcement Learning (Q-DRL) algorithm. The research will provide a new level of understanding the behavior of economic agents with respect to preferences, captured by quantiles, without assuming a specific utility function or distribution of returns. Furthermore, we are challenging the traditional diversification methods as they proved to be insufficient due to heightened correlations and similar risk features between asset classes, and rather the research delves into risk factor investing as a solution and portfolio optimization based on them.

Keywords: Portfolio Management, Quantile Deep Reinforcement Learning, Factor investing, Deep-Learning, Advantage-Actor-Critic

1. Introduction

Modern portfolio theory (MPT), established by Markowitz (1952), assumes that investors will take higher risk, if it is compensated by higher expected returns and the risk-return trade-off depends on individual risk aversion coefficients. MPT assumes that asset returns are normally distributed and the correlation between them is constant. Consequently, it cannot capture statistical features of risk and return which are often highly skewed. Another prominent classical method is the Capital Asset Pricing Model (CAPM), which introduced the concept of the risk-free rate and the market risk premium. It assumes that asset returns are linearly related to their market risk. CAPM's risk budgeting principle allocates investments based on their sensitivity to systematic risk (beta). While CAPM provides valuable insights, it has been criticized for its simplifying assumptions and sensitivity to market conditions. Furthermore, MPT and CAPM only give a solution to a single-period optimization, which raised the need of dynamic portfolio allocation models (DPA) to capture time-varying risks and opportunities. The benefits of dynamic models are proved by Ang and Bekaert (2004), who showed that over bear markets investors tend to switch to cash, which is advantageous due to increasing interest rates. Similarly, Guidolin and Timmermann (2007) indicated that the optimal portfolio is dynamic and changing according to regimes (crash, slow growth, bull and recovery states). However, to switch the allocation conditionally on market regimes, we need to forecast these states or adapting to them immediately, which is difficult due to imperfect markets. Wang and Aste (2023) introduced a pioneer DPA algorithm , which identifies the inherent market states and forecast future state. Khedmati and Azin (2020) developed a multiple step market states clustering method to dynamically update the optimal portfolio at the beginning of each period. Finally, Khan and Mehlawat (2022) developed a two-phased DPA method by clustering time series with technical indicators considering the risk aversion of investors and at the second step they optimized the portfolio weights.
To accurately reflect market dynamics, it's essential to consider the current state or regime of the market. The dynamics of changes can be captured by Markov chain switching models (Hematizadeh et al. (2022), Fons et al. (2021), Nystrup et al. (2018). However, it's important to note that these approaches often have limitations due to their rigid assumptions about the conditional return distribution, transition probabilities, and the complexity arising from a high-dimensional state space.
To overcome these challenges, reinforcement learning (RL) can serve as a solution. RL has the capacity to handle high dimensional spaces, approximate any distributions, learn online with trial and error while considering exploration/exploitation trade-off. The main features of reinforcement learning according to Sutton and Barto (2018) is that the learner is not told what to do but discover and the decisions made are effecting not just the next state and reward but all subsequent rewards. The main goal of any RL method is to find an optimal policy (π^*), which is maximizing the expected cumulative reward (i.e sharp ratio, portfolio value). We can differentiate 1) value-based methods, where we estimate the value function (state (V^*) or state-action value (Q^*), which will lead us to the optimal policy π^*. The policy is implicitly derived from these estimated values, since the policy itself is not trained, consequently, we need to specify a behaviour how to choose actions by using the estimated values. For example in Q-learning, ϵ-greedy policy is used, which chooses the highest estimated value with probability $1 - \epsilon \in [0, 1]$ and a random action with probability ϵ. 2) Policy-based methods on the other hand directly learn the optimal policy π^*, without having the value function. The input of the policy is the state (i.e accumulated information of the market) and the output is a probability distribution over actions in case of stochastic policy. However, value based methods may experience high bias due to the non-stationary nature of MDP transitions in financial markets, where the environment's dynamics constantly change and policy-based methods often also leads to high variance (Yang, 2023). To address these challenges, Actor-Critic methods, such as Advanced-Actor-Critic, have emerged, seeking to balance the strengths and mitigate the weaknesses of both approaches. The model contains two neural networks, the Actor that controls the agent behavior by choosing actions and the Critic that gives a feedback to the actor how good the taken action is. The

extension of the vanilla Actor-Critic method by Google DeepMind (2015), capable of handling continuous, high-dimensional state spaces, is a significant milestone. Following this Liu, Xiong, et al. (2018) explored DDPG (Deep Deterministic Policy Gradients) in portfolio optimization (PO) problems and outperformed the min-variace and DJI index.

Many of the RL models related to PO tried to capture the effective state representation. Ye et al. (2020) incorporated price movement predictions in the state representation, which lead more robust results in uncertain environments. Similarly, Yang (2023) introduced a Task-Context Mutual Actor–Critic algorithm to represent the state but also in a global dynamic way by using attention-based mechanism. Lucarelli and Borrotti (2020) proposed a multi-agent model where each local agent in the system employs a specialized deep Q-learning approach to contribute to a global reward function. This function, managed by a global agent, dynamically updates the information for each local agent.

Previous researches have explored the application of RL in financial markets, primarily focusing on maximizing expected cumulative rewards, using not economical data and neglecting the significant risks of potential tail events and risk appetite of investors. An effective approach to address this gap is to implement distributional reinforcement learning methods that provides risk-sensitive policies. Wang and KU (2022) proposed a Hierarchical DDP algorithm for portfolio management, integrating DDPG with a hierarchical structure. The high-level policy operates at an abstract layer, assigning sub-tasks to a lower-level policy that executes specific targets. The model incorporates a parametric Conditional Value-at-Risk (CVaR), allowing the high-level policy to adjust actions based on portfolio risk. This structure maps the state to different actions by maximizing the α percentile expectation based on different values of risk parameter α. Similarly to Tang et al. (2019) their model is based on the assumption that the return distribution is Gaussian, which lead to a closed-form of solution.

An equally important aspect, with state representation and model selection, is the choice of which stocks to optimize, as different stocks may be subject to the same risk factors. This consideration naturally guides us towards factor investment strategies. In the context of investment, a factor refers to any attribute that can significantly influence an asset's risk-return trade-off. In general, factors are essentially key determinants used to understand and predict the performance and behaviour of financial assets. Numerous studies in the empirical asset pricing literature showed that specific factors, such as the Fama-French (2015) factors, are providing higher risk-adjusted returns. Furthermore, Nazaire et al. (2021) discovered that diversifying across various factors in investment portfolios significantly enhances downside protection. Factor investing is also a key for portfolio management for institutional investors (Dopfel and Lester (2018)). The authors noted that the development of advanced beta, comprising complex multifactor investments, likely hasn't led to portfolios with known or ideally balanced cumulative exposures, which means the ideal mix of factors in the portfolio is unknown. Staden et al. (n.d.) developed a neural network approach to find optimal dynamic factor investing strategy by using 2 objectives, mean-Cvar and one-sided quadratic target. The study raises a crucial issue with optimal investment strategies derived from training and testing data sets, notably the lack of substantial factor diversification. They argued that this limitation mainly stems from employing highly correlated, long-only equity factor indices or ETFs. Additionally, despite promising in-sample investment results from optimal factor investing strategies, these strategies often show a lack of meaningful diversification among the factors themselves, suggesting a potential area for improvement in investment strategy development. The previous paper formulated factor investing as a stochastic optimal control problem, which could be solved by RL as it was done by Andre and Coqueret (2020). The authors address a significant issue encountered with conventional neural networks in portfolio management: the inability to convert network outputs directly into portfolio weights. To overcome this, they proposed the use of Dirichlet distribution and documented that the portfolios shaped by reinforcement learning strategies are closely aligned with the equally-weighted (1/N) allocation.

Our research intends to fill two gaps in the literature. To allocate assets across factors is unquestionably important but the literature haven't concluded how different factors effect the distribution of returns and

they documented close to equally weighted factor portfolios, which seems unrealistic. To solve this dynamic optimization, distributional RL is a natural choice. Our goal is to create a model that prioritizes quantile preferences in a straightforward and accessible manner. This model is designed to efficiently explore without relying on specific assumptions about return distributions or utility functions.
The paper is organized as follows. Section 2 summarizes the basic concepts of both Actor-Critic and the closely connected distributional RL models. Section 3 describes our data set. In Section 4, we introduce our methodology. Section 5 discuses our model results and Section 6 concludes.

2. Advantage-Actor-Critic and Distributional Reinforcement Learning

2.1 Advantage-Actor-Critic

This section intends to state the main ideas and equations of the closely related models. Our model is based on Advantage-Actor-Critic method and its closely related distributional version. The vanilla Actor-Critic method has two main components: an Actor and a Critic. The Actor is responsible for choosing actions based on the current policy, formulated in terms of a probability distribution over actions conditionally the current state. The Critic evaluates the chosen action by computing the value function, which estimates the expected reward from the current state onwards. The Actor-Critic method aims to optimize the policy by adjusting the Actor's parameters based on the feedback from the Critic. By iteratively updating both the Actor and the Critic, the method seeks to find an optimal policy that maximizes the cumulative reward over time.

The policy gradient of the Actor is

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t\right] \tag{1}$$

where, R_t is the cumulative reward. To reduce the variance of the policy, we can subtract a regularization term, which is called the baseline ($V_\phi(s_t)$ in our case).

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) A(s_t, a_t)\right] \tag{2}$$

where $A(s_t, a_t)$ is the so called advantage value, which is:

$$A(s_t, a_t) = Q_\phi(s_t, a_t) - V_\phi(s_t) = r_t + \gamma \cdot V_\phi(s_{t+1}) - V_\phi(s_t) \quad \text{(via Bellman optimality)} \tag{3}$$

Intuitively, by subtracting the cumulative reward with a baseline leads to reduced gradients, thereby resulting in smaller, more stable updates. The introduced advantage function indicates, that how advantageous a particular action is, in comparison to an average action, in a given state. Consequently, the neural network will learn more effectively those actions, which result in higher cumulative rewards.
The Critic model is responsible for estimating the value function $V(s)$, which is parameterized with ϕ. The objective function of the Critic is

$$J(\theta) = r_t + \gamma \cdot V_\phi(s_{t+1}) - V_\phi(s_t) \tag{4}$$

2.2 Distributional RL

This section is focusing on the main ideas of the Distributional Reinforcement Learning, specifically Distributional Soft Actor Critic (DSAC) by Ma, Xiaoteng, et al. (2020).

2.3 DSAC

Our research is most closely aligned with the distributional version of the Soft-Actor-Critic method (SAC), as described by Haarnoja et al. (2018). SAC optimizes a policy by considering both the reward and the randomness of the action (entropy) leading to improved exploration and stability. The model leverages an off-policy approach in its Actor-Critic framework, meaning it learns from both current and past experiences, using a stochastic policy. This structure is particularly effective for continuous problems, ensuring robust learning by utilizing a broader range of data for updates, rather than relying solely on the latest experiences. The objective function of SAC is

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left[R(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right] \right] \tag{5}$$

and the Soft-Bellman operator is defined as

$$T_S^\pi Q(s,a) := \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{P,\pi} \left[Q(s',a') - \alpha \log \pi(a'|s') \right] \tag{6}$$

SAC involves an iterative process where soft-policy evaluation is followed by soft-policy improvement. The updating process is done by minimizing Kullback-Leibler divergence between the policy distribution and exponential form of soft action-value function.
SAC takes the action space randomness into account via the optimization procedure but it neglects the random behaviour of the reward distribution, which is done by DSAC.
Let $\mathcal{Z}^\pi : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$ denote the soft action-value distribution of a policy $\pi \in \Pi$, which is defined by the authors as

$$Z^\pi(s,a) :\overset{D}{=} \sum_{t=0}^{\infty} \gamma^t \left[R(s_t, a_t) - \alpha \log \pi(a_{t+1}|s_{t+1}) \right] \mid a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t), s_0 = s, a_0 = a. \tag{7}$$

and the distributional Bellman, which consist both the randomness of the reward and the actor, becomes

$$T_{DS}^\pi Z(s,a) :\overset{D}{=} R(s,a) + \gamma \left[Z(s',a') - \alpha \log \pi(a'|s') \right] \mid s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s'). \tag{8}$$

By repeatedly applying the similar updating process as in the SAC, the policy will finally converge to the optimum.
In DSAC the Critic, $Z_\tau(s,a;\theta)$, assesses the value of actions in states, while the Actor, $\pi(a|s;\phi)$, chooses stochastically the actions based on the state. To train the Critic we need to define $Z_\tau := F_Z^{-1}(\tau)$, where F_Z^{-1} is the quantile function. The action-value distribution is approximated with quantile fractions, which satisfies $\tau_0 = 0, \tau_N = 1, \tau_i < \tau_j, \forall i < j,$ and $\tau_i \in [0,1], i = 0, \ldots, N$ and the authors denoted $\hat{\tau}_i = \frac{\tau_i + \tau_{i+1}}{2}$. The $Z_\tau(s,a;\theta)$ quantile regression is optimized through the minimization of the weighted pairwise Huber regression loss across various quantile fractions. It can be shown that the original maximization problem can be expressed as

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}} \left[\log \pi \left(f(s_t, \epsilon_t; \phi) | s_t \right) - Q \left(s_t, f(s_t, \epsilon_t; \phi); \theta \right) \right] \tag{9}$$

, where ϵ is Gaussian noise vector used by the reparametrized policy neural network f(s, ϵ, ϕ) and $\mathcal{D}$ is the transitions replay buffer.

To introduce risk measures the authors define $\Psi : \mathcal{Z} \to \mathbb{R}$, which is mapping from the value distribution to a real number. The risk soft action-value, therefore $Q_r(s,a) = \Psi[\mathcal{Z}(s,a)]$, where $\mathcal{Z}(s,a)$ is from equation 7. The natural risk measures are the VAR, mean variance and distorted expectation.
DSAC is an advanced extension of SAC by introducing quantile loss function to train the Critic. On the other hand the risk measures are difficult to understand as they objective contains exploration, which deviates from the standard risk measures. Furthermore, while DSAC's off-policy learning from past experiences broadens its learning scope, it sacrifices the understanding of temporal sequences in decision-making.
To help to understand, how a risk sensitive decision maker works over factor investing, we would like to develop a quantile preference decision maker, which is easy to understand and train.

3. Data

Numerous studies in empirical asset pricing literature have demonstrated that certain factors offer higher risk-adjusted returns. However, even as a growing number of new factors (factor zoo) emerge aiming to clarify the cross-section of expected returns, Liu and Zhu (2016) developed an estimation strategy, finding that most of these research findings lack statistical significance. In a similar vein, Feng et al. (2020) explored the importance of newly identified factors in finance, discovering that while many don't contribute additional insights beyond existing knowledge, only a few possess substantial explanatory power. Consequently, we choose the most researched Fama-French factors and added some easily interpretable new ones. Factor data is represented by ETFs from BlackRock, provided by Thomson Reuters (TR). The definition and some characteristics of the factor ETFs can be seen in Table 2. These ETF products are considered to be relatively new, consequently our data set's range is from 2015-12-14 to 2023-12-08. The correlation between assets is surprisingly high (Figure 2). Most probably because these ETFs are focusing on US stocks and the products are not necessary exclusive. Figure 1 shows the closing prices of the ETFs. Abrupt structural breaks and price corrections, particularly evident at the end of 2019 and 2021, can disrupt the learning process in the Critic network potentially leading to inaccurate predictions, also, the Actor will be less confident, which actions to take.

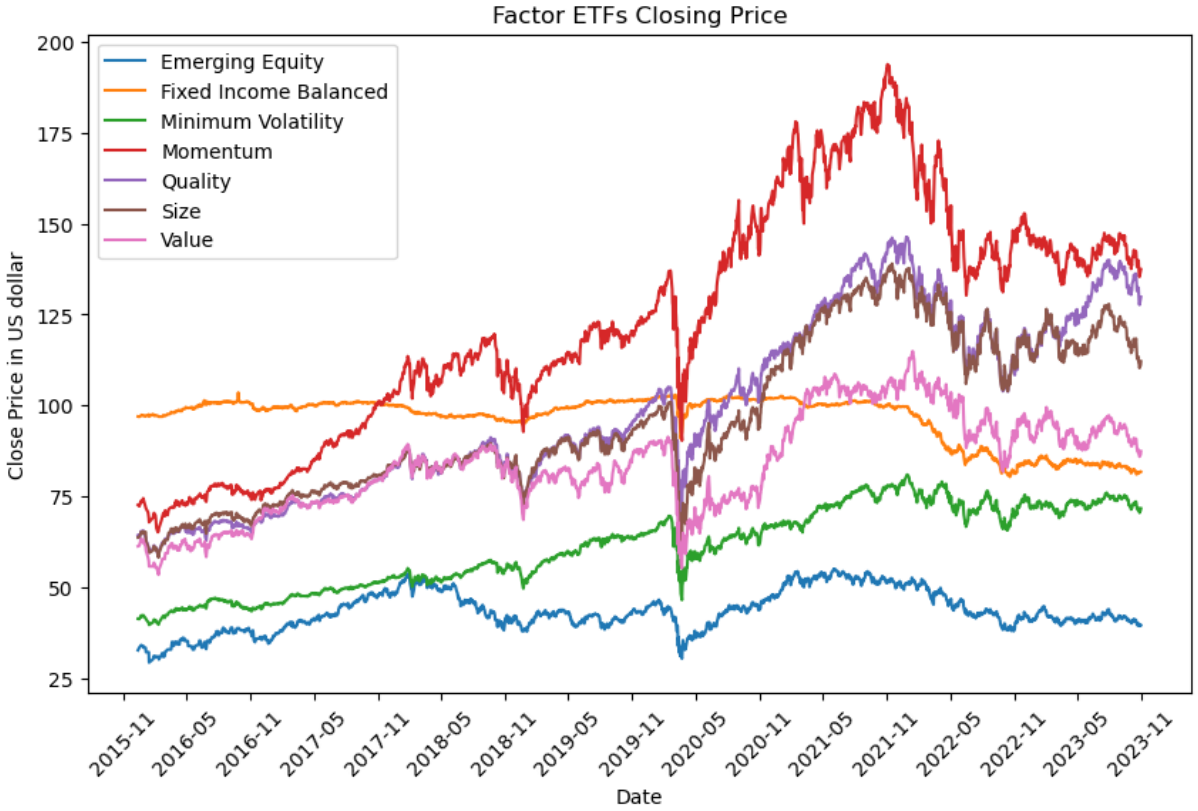

Figure 1: ETFs Closing prices

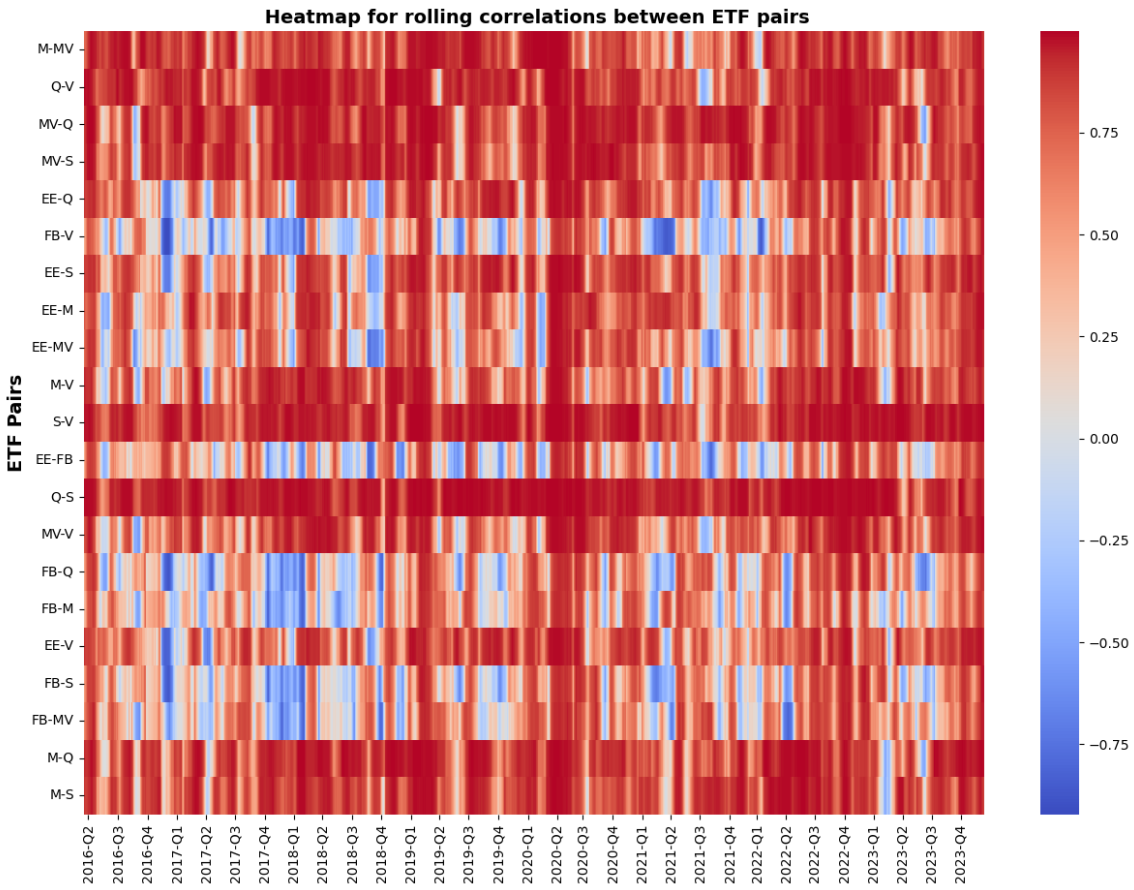

Figure 2: Rolling Correlation Heatmap for ETFs
Note: Emerging Equity: 'EE', Fixed Income Balanced:'FB', Minimum Volatility:'MV', Momentum:'M', Quality:'Q', Size:'S', Value:'V'

4. Methodology

4.1 Embedding PO to RL

We utilize daily reallocations to dynamically adjust the investment distribution among various factor ETFs. By choosing daily reallocations over less frequent (e.g., weekly) adjustments, we avoid introducing an additional hyperparameter, thus simplifying the model's interpretability and most probably causing higher variance in the daily asset weights.

The PO problem will be solved by using RL, which enables the agent to dynamically adapt and refine strategies through continuous interactions with the environment, learning from trial and error. The portfolio environment, which is responsible for stock market simulation, is a modified version of the FinRL environment from Liu et al.(2018).
In general RL environment has three main components:

- **Action Space:** PO involves a continuous action space, implying that the agent allocates weights across assets in a continuous range from 0 to 1. The agent will execute the allocation task for each ETFs in each trading day. After this allocation, the weights are normalized using the Softmax

function to ensure their total equals to 1. No short-selling is allowed, no transaction costs and cash reserve. The action space has size 7, which is the number of factors.

- **State Space:** The state space in portfolio optimization encompasses all the features that the agent is expected to observe at each time step. This includes various market indicators, historical data, and relevant financial metrics, providing the agent with a comprehensive view of the market conditions to inform its decision-making process. We chose explanatory variables mainly focused on specific factors, but some are versatile enough to provide insights into several factors. The data is coming from TS except for VIX and 10-Year-Treasury. The provided variables are in daily bases but Business Cycle Indicators are monthly. The structure of the state space, represented as a matrix of (33,7), integrates 19 core features alongside an additional 14 features derived from two sets of covariance matrices. The dimension '7' corresponds to the number of Factor ETFs considered.

  - **Breadth Indicators:**[1] Breadth indicators are statistical measures used in finance to assess the number of stocks advancing versus declining. They provide insights into the overall health and direction of a market. Our indicators reveal the percentage of S&P 500 stocks trading above their 50-day and 200-day simple moving averages of closing prices. This provides a clear view of the short-term and long-term market trends, offering valuable insights into the overall market momentum.

  - **Business Cycle Indicators:**[2] Our initial data, provided on a monthly basis, was resampled to a daily resolution. To bridge the gaps created by this resampling, first, we shifted ahead the monthly data and we applied third order polynomial interpolation to fill in the missing days. This shift is essential because, for instance, data from February, which becomes available at the month's end, isn't accessible at the start of February. This technique ensures that the daily dataset remains consistent, continuous and avoid information leakage.

    * *US Leading Economic Index:* This variable serves as a forecaster, anticipating shifts in the business cycle about 7 months ahead. It has 10 components such as average weekly hours in manufacturing, average weekly initial claims for unemployment insurance etc
    * *US Coincident Economic Index:* This index provides information of the current state of the economy. The components are payroll employment, personal income less transfer payments, manufacturing and trade sales, and industrial production.
    * *US Lagging Economic Index:* It serves to confirm economic turning points and the strength of new trends. The index includes seven components like unemployment, unsold inventories.
    * *US Leading Credit Index:* It consist bearish and bullish investor sentiments, Treasury-Bill yield spread, Swap-spread etc. It considered to be a forward looking index.

  - **GOLD, USD/EURO:** The USD/EUR exchange rate significantly influences international trade and economic activities, impacting the global market sentiment. GOLD is negatively correlates with interest rate, which can be an anticipator of stock movements and Fixed Income ETF.

  - **S&P 500 calculated P/E:** S&P 500 serves as a benchmark about the sentiment and strengths of the overall economy, consequently it may be a general explanatory variable for each factors.

[1]Note that we took the recent SPY500 components' data until the beginning of our examination period, which means it is not the best representation as the index components are dynamic.

[2]Description: Conference Board,Investopedia

- **MSCI Emerging Markets Index (EMI):**[3] This index targets large and mid-cap stocks in 24 Emerging Markets, focusing on those with value-style characteristics. Index inclusion is based on three key variables: the ratio of book value to price, the ratio of 12-month forward earnings to price, and the dividend yield. Our Emerging Factor ETF is very similar to the MSCI EMI index, but the latter is much broader.
- **10-Year Treasury Constant Maturity Rate:**[4] Its movements can influence various market segments, including equities and bonds, consequently it may explain the variance of Fixed Income and Minimum Volatility ETFs.
- **VIX:**[5] Its movements can influence various market segments, including equities and bonds, consequently it may explain the variance of Fixed Income and Minimum Volatility ETFs.
- **Others:**: We also include the rolling percentage changes (window size 5, 10) of the ETFs' closing price, simple moving averages (SMA) with window size 5 and 20, rolling correlation matrices with size 128 and 252 trading days and the bollinger bands, which is the 2 times standard deviation of the closing prices from the SMA with time window 20.

- **Reward:** In RL reward has a crucial impact on the learning process. It is a signal provided to the learning agent from its environment, indicating the success or failure of its actions. This reward guides the agent in learning optimal strategies by reinforcing behaviors that lead to higher rewards/outcomes. We already take risk into account by using the quantile set up, consequently it is logical to use the scaled portfolio returns as a reward. Scaling is important for numerical purposes.

$$\text{Portfolio Return} = \sum \left(\left(\frac{\text{Current Price}}{\text{Previous Price}} \right) - 1 \right) \times \text{Weights} \tag{10}$$

$$\text{Reward} = \text{Portfolio Value} \times (1 + \text{Portfolio Return})\text{*Scaling Factor} \tag{11}$$

4.2 Model implementation: Q-A2C

The model we have developed is an on-policy distributional Advantage (TD)-Actor-Critic framework, where the Actor is stochastic via Gaussian policy, with parameter μ and σ, and the distributional effect is captured by the Critic via quantile loss and monotonicity penalty. The model's Critic network is parameterized by θ and it outputs quantile values for each τ-s. Similar to Dabney et al.(2018) and Ma, Xiaoteng, et al.(2020) notation, the distributional Bellman becomes

$$T_D^\pi Z_\theta^\pi(s) :\overset{D}{=} R(s,a) + \gamma Z_\theta^\pi(s') \mid s' \sim P(\cdot \mid s,a), a \sim \mathcal{N}(\mu_\phi(s), \sigma_\phi(s)) \tag{12}$$

where R(s, a) is the immediate reward for state 's' after taking action 'a', γ is the discount factor, $s' \sim P(\cdot \mid s,a)$ denotes the probability of transitioning to the next state, and the last part signifies the action taken by the Actor network parameterized by ϕ. The temporal difference is

$$\delta_{t,i} = R(s_t, a_t) + \gamma Q_\theta(s_{t+1}, \tau_i) - Q_\theta(s_t, \tau_i) \tag{13}$$

and the corresponding Critic loss function becomes

$$\text{loss}_\tau = |\tau - \mathbb{I}\{\delta_{t,i} < 0\}| \times |\delta_{t,i}| \tag{14}$$

$$\mathcal{L}(\theta) = \frac{1}{|T|} \sum_{\tau_i \in T} \text{loss}_{\tau_i} + \lambda \cdot \frac{1}{T-1} \sum_{i=1}^{T-1} \max(0, Q_\theta(s_t, \tau_i) - Q_\theta(s_t, \tau_{i+1}) + \epsilon) \tag{15}$$

[3]Description: MSCI Emerging Markets Index
[4]Data Source: 10-Year Treasury Constant Maturity
[5]Data Source: VIX

where the last term of the Critic loss is the quantile penalty to ensure monotonicity of the estimated quantiles.
We've opted for a Gaussian policy for the Actor, primarily because it's well-suited to address continuous PO problems. The Gaussian policy's high entropy characteristic is especially advantageous as it facilitates exploration. To further encourage exploration we added an entropy regularization term for the Actor loss function.

$$\mathcal{L}(\phi) = -\left[\sum_i \log \pi_\phi(a_i|s)\right] \cdot \delta_\tau - \lambda \cdot H(\pi_\phi(s)) \tag{16}$$

By integrating the quantile-specific TD error (Equation 13) with overall action log-probabilities, represented as portfolio weights, the model aims to uniquely tune returns within this selected quantile, all under the Critic's guidance. More precisely, the Actor enhances the likelihood of actions that yield returns exceeding the Critic's expectations, indicated by a positive TD error, while the Critic continuously refines its evaluation of potential outcomes, providing updated and more precise quantile-based feedback. Furthermore, incorporating entropy into the loss function ensures a well-balanced approach between exploration and exploitation and make the model more interpretable. Note that the interpreation and the learning procedure is the same for using advantage and the traditional quantile specific loss:

$$\text{loss} = \sum (\log \pi_\phi(a|s) \cdot |\tau - \mathbb{I}\{\delta_{t,i} < 0\}| \cdot |\delta_{t,i}|) \tag{17}$$

The Actor is designed to maximize returns by influencing the quantile-specific error (TD error) to be positive, thereby making the quantile loss zero. This approach aligns with TensorFlow's optimization framework, which inherently transforms maximization problems into minimization tasks. On the other hand, the indicator function would make the training more complex and we wouldn't be able to exploit the variance decreasing property like in the advantage case.

4.3 Implementation and Neural Networks Structure

The algorithm is summarized in Algorithm (1) and the parameters can be found in Table 1. To facilitate a more direct comparison, we employed uniform parameters across all quantile estimations. However, this approach brings challenges, particularly in a financial context where distinct segments of the return distribution exhibit varying risk-return profiles. Uniform parameters may not accurately capture the unique dynamics of each quantile, potentially leading to suboptimal performance.
Our model is using gradient descent algorithm to find optimal solution. The gradients, which are central to the learning process, can vanish or explode if the data is not properly scaled. Furthermore, when input features have vastly different scales, some weights might update faster than others, leading to erratic learning patterns. Scaling ensures all input features contribute approximately equally to the learning process, making training more stable and helps to converge. We may differentiate 3 types of scaling/normalization. Batch normalization, normalizes the inputs across the batch dimension, layer normalization[6], on the other hand, normalizes the activations of each layer independently across all features (Ba et al.(2016)) and lastly we can do feature scaling for each feature independently. In the originally developed default model, we prioritize interpretability by minimizing the number of parameters. To achieve this, we avoid batch training and opt for standard scaling at each step. For a state with a shape of (33, 7), we perform standardization across all features, which is similar approach to layer normalization. Introducing layer normalization would entail additional affine transform parameters, potentially complicating the model's interpretability and slowing down the training process. Note that if we would use the only traditional machine learning standard scaling independently for each feature per ETF, the model would not work, we

[6] Layer Norm Implementation: Pytorch

would get infinite values for the sigmas. For future research, we may try MinMax scaling, Convolutional layers and layer normalization.

The Actor network is initialized to process a state space reflecting numerous market features and stock dimensions. It employs a deep neural architecture, comprising two dense layers with 64 units each, using leaky ReLU activations and HeNormal kernel initialization, which is particularly effective for this activation function. The network includes high L2 regularization, decaying learning rate and small dropout for robustness against overfitting. In general, decaying learning rate schedules gradually reduce the step size during training, allowing the optimization process to make smaller adjustments as it approaches a local minimum. L2 regularization penalizes large weights, encouraging simpler models and preventing overfitting, similar to dropout, which can help avoid getting stuck in complex, high-dimensional local minima.

After flattening the processed input, the network outputs two crucial parameters: '$\mu_\phi(s))$', denoting the mean of the proposed action distribution, and '$\log \sigma_\phi(s)$', indicating the logarithm of the standard deviation. Using '$\log \sigma_\phi(s)$' instead of standard deviations enhances the stability and efficiency of the model. Log standard deviations can range freely across $(-\infty, \infty)$, making them easier to train since they don't require the enforcement of non-negativity constraints like standard deviations. The actual standard deviations will be derived by exponentiating the '$\log \sigma_\phi(s)$', ensuring no loss of information. The Actor network samples actions from a multivariate Gaussian distribution defined by using the estimated previously defined parameters. The sampling process makes our network continuous and stochastic, which is essential for both exploration and imitation of the portfolio choice problem. To enhance exploration in our approach, we avoid constraining sigma, and we do not employ gradient clipping for the Actor, even if the gradient norm surpasses a specified threshold. In case of Critic network we normalize the gradients by using norm clip method by TensorFlow.[7]

The Critic network differs from the Actor in several key aspects. While it also processes the state inputs and employs a similar structure of two dense layers, its focus and output are distinct. The Critic utilizes a linear activation for its output layer, contrasting with the Actor's Gaussian distribution outputs. Furthermore, the Critic's output size is set to the number of tau levels, designed to provide valuations across various quantiles of potential returns. Note that the inherent randomness of the model, primarily stemming from the initial random initialization of network weights and the stochastic nature of the action sampling process, introduces variability in its behavior and outputs. This randomness can lead to challenges in achieving consistent results across different runs, potentially impacting the replicability and the training process. We run the algorith both for the training and testing over 30 episodes and analyze the last episode output.

[7] Norm Clip TensorFlow

Table 1: Parameters for the Q-A2C Model

Parameter	**Value**
Rho (soft update parameter)	0.001
Tau Levels	10
Entropy Regularization	0.001
Gamma(Discount Factor)	0.99
Critic Learning Rate (Start)	0.00001
Critic Learning Rate (End)	0.000001
Actor Learning Rate (Start)	0.000005
Actor Learning Rate (End)	0.0000005
Episodes	30
Learning Tau	0.9, 0.5 and 0.2
Polynomial Decay rate	2
Monotonicity Penalty	1.4
Critic Gradient Clipping Norm	1
L2	0.01
Dropout	0.1

Algorithm 1 Q-A2C Training Process

Require: Initial parameters for state, stock dimension, networks (Actor and Critic) and number of episodes M

Note: y is the target for training the Critic network. It is computed as the observed reward plus the discounted future value estimated by the Target Critic Network.

Randomly initialize Actor Network: Actor(s, ϕ)
Randomly initialize Critic Network: Critic(s, θ)
Initialize Target Critic Network: Critic Target$(s, \bar{\theta}) \leftarrow$ Critic(s, θ) using soft update
for episode = 1 to M **do**
    Initialize state s
    Reset terminal flag and total reward
    while not terminal **do**
        Sample action a using Actor Network: $a \sim \mathcal{N}(\mu_\phi(s), \sigma_\phi(s))$
        Execute action a, observe reward r, next state s', and terminal flag
        Compute target y for non-terminal or set $y = r$ if terminal
        Perform learning step with $(s, s', a, r, \text{terminal})$
        Update Critic: $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta \mathcal{L}(\theta)$
        Update Actor: $\phi \leftarrow \phi + \alpha_\phi \nabla_\phi \mathcal{L}(\phi)$
        Soft Update for Critic Target: $\bar{\theta} \leftarrow \rho\theta + (1 - \rho)\bar{\theta}$
        Update state to s' and accumulate reward
        if Terminal **then**
            Save losses, training weights
        end if
    end while
    Output episode reward and other episode-level metrics
end for

5. Results

Reinforcement learning is challenging to understand due to the complexity of learning the optimal behavior through trial and error in a dynamic and not stationary environment. Additionally, our Q-A2C framework gives additional complexity because 3 neural networks work together to solve a non-closed form solution problem.
We know the model is learning if 1) we have intuitive result, 2) the loss functions are decreasing over time, 3) the agent experienced exploration and the entropy is in a specific not extreme range and optimally decreasing over episodes, 4) the cumulative reward is increasing, 5) the monotonicity penalty is in a specific range and optimally decreasing. We'll analyze outcomes across Tau levels (2, 5, and 9), with a focus on loss functions, as they guide the optimization by offering performance feedback, aiding adjustments for accuracy and convergence. Additionally, we'll examine cumulative rewards to further understand model effectiveness, providing insight into the agent's long-term performance and its ability to maximize rewards in a specific quantile over time, which is the main task of the Actor network. We also compare the maximum drawdown for each tau, which serves as a measure of the largest loss experienced by the investment strategy within the specified time frame. This comparison helps assess the risk exposure and downside potential across different quantiles.

5.1 Tau 2: Risk Averse

Tau 2 is the representation of a risk averse investor. After reviewing Figure 2, we may suspect that the agent will choose a mix of Emerging Equity, Fixed income or Minimum Volatility ETFs. It is also possible, that the agent focusing on the lowest volatility ETFs and a high volatility one to maximize the quantile specific risk-return trade-off. Figure 5 and Figure 6 shows the train-test results. In both sets the agent chose the lowest beta and standard deviation assets and neglecting the riskier ETFs such as Momentum or Size. We grouped the training episodes' outcomes into 3 groups and plot the results. It is visible, in average, there is policy improvement through the decreasing loss pattern (Figure 16,19,17,18). At step 800, which corresponds to the end of 2019 period, there is a structural break, which causing high volatility in the loss functions. The entropy has significantly changed in the last group comparing to the first, on the other hand the exploration still remained. Note that the agent Actor's loss is a decreasing function of the entropy (and in TensorFlow we do minimization), which means by construction we support exploration. On the other hand, the agent decreases the entropy over time, which means it is more confident what actions to take. At the Critic, the sudden change observed in the Actor, Critic and TD loss patterns towards the end of episodes can be attributed to the terminal stage, where the agent receives only a reward without discounting future values, as specified in Algorithm 1. As a result, the agent's behavior may undergo abrupt shifts, prioritizing short-term gains over long-term strategies.
The cumulative sum of returns (rewards) is relatively low over the training period (Figure 3). For a risk aversion agent is an expected behavior if the maximum drawdown (4) is smaller than for other taus. The agent's primary objective to earn higher discounted rewards in a specific quantile consistently over time, reflecting its ability to make optimal decisions throughout the training process. In case of Tau 2 it is satisfied.

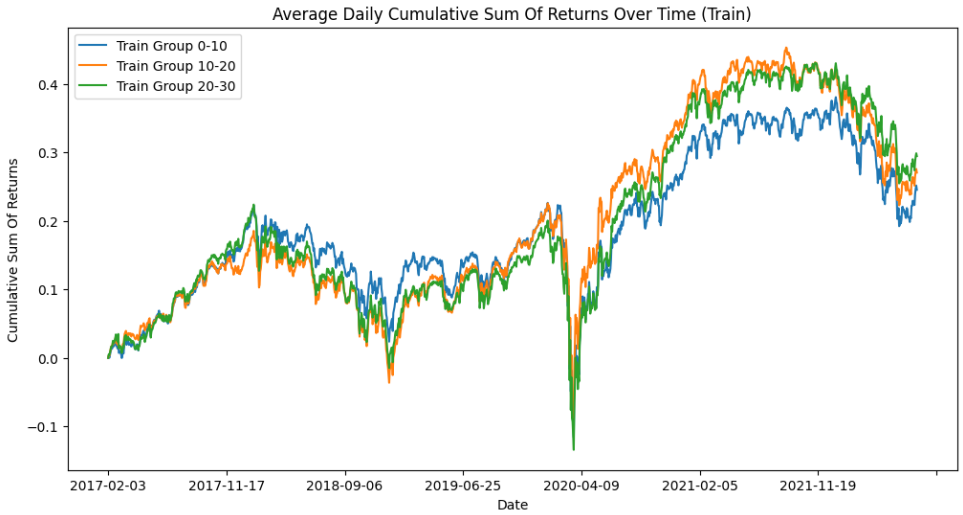

Figure 3: Tau 2 cumulative sum of returns/rewards: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

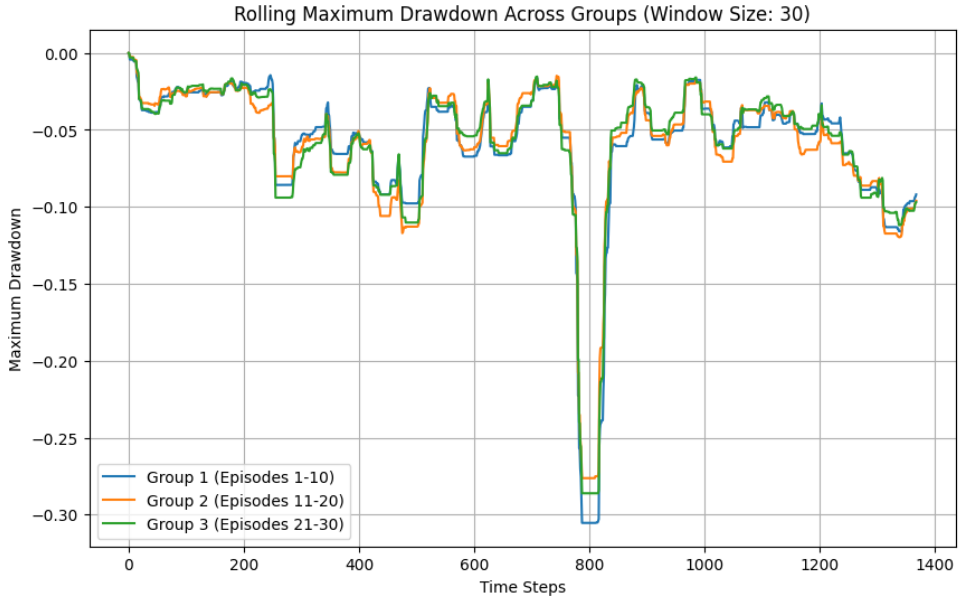

Figure 4: Tau 2 maximum drawdown: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

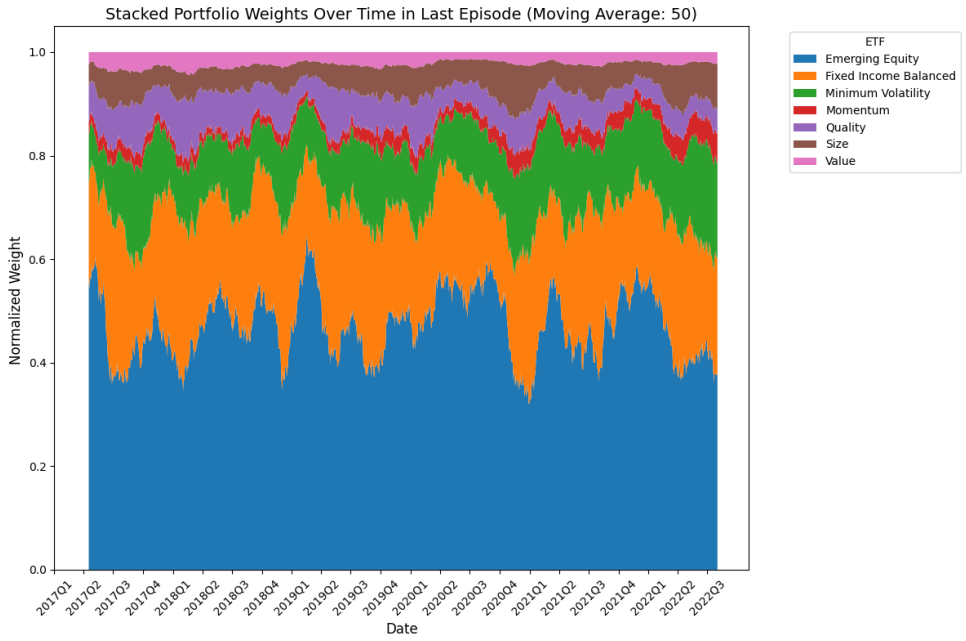

Figure 5: Tau 2 asset allocation: Training set

Note: We used 50-days moving average on portfolio weights while plotting the graph. Both in the train and test set we used 30 episodes and plotted the last episodes outcome.

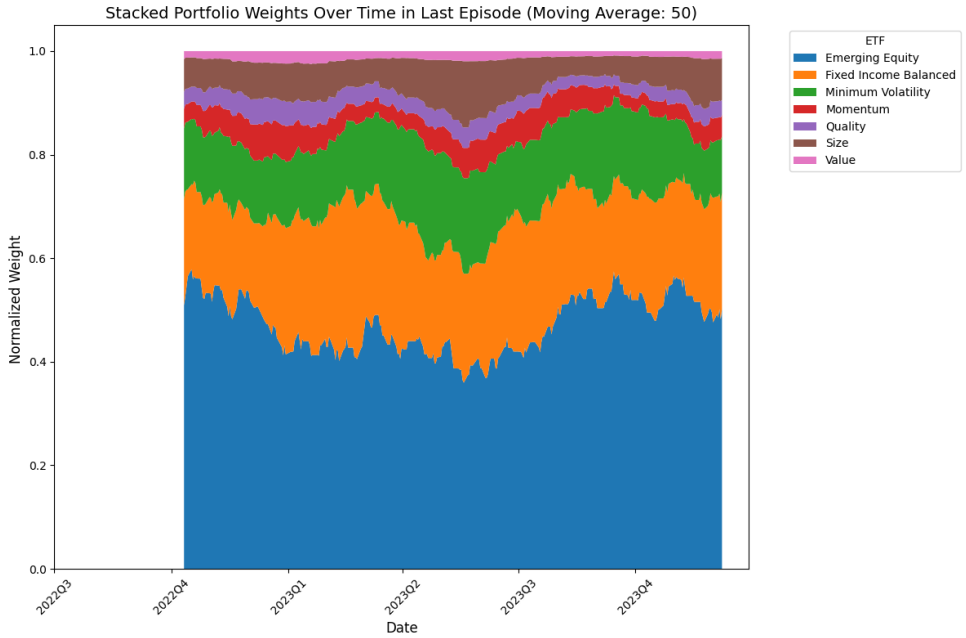

Figure 6: Tau 2 asset allocation: Test set

Note: We used 50-days moving average on portfolio weights while plotting the graph. Both in the train and test set we used 30 episodes and plotted the last episodes outcome.

5.2 Tau 5: Risk Neutral

Tau 5 asset allocation (Figure 9, Figure 10) is different from Tau 2. The agent allocates more in risky-return profile ETFs like Momentum or Quality. The training outcome shape (Figure 20, Figure 21, Figure 23, Figure 22) is similar to Tau 2. On the other hand if we closely look and compare the loss functions magnitude, we may deduct that in case of Tau 5 the losses are higher, which means it is easier to to find a less-risky strategy. One economical reasoning is the high correlation between assets and the general

high volatility in the factor etfs.

The cumulative sum of rewards/returns (Figure 7) are increasing over time and episodes. The maximum drawdown (Figure 8) is higher than for Tau 2, which suggest that the agent indeed managed to capture the left-tail of the returns.

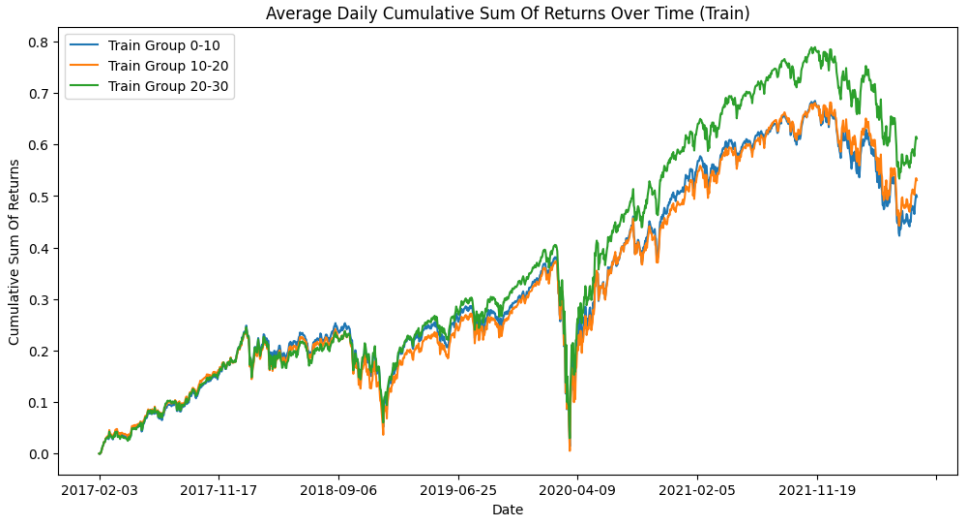

Figure 7: Tau 5 cumulative sum of returns/rewards: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

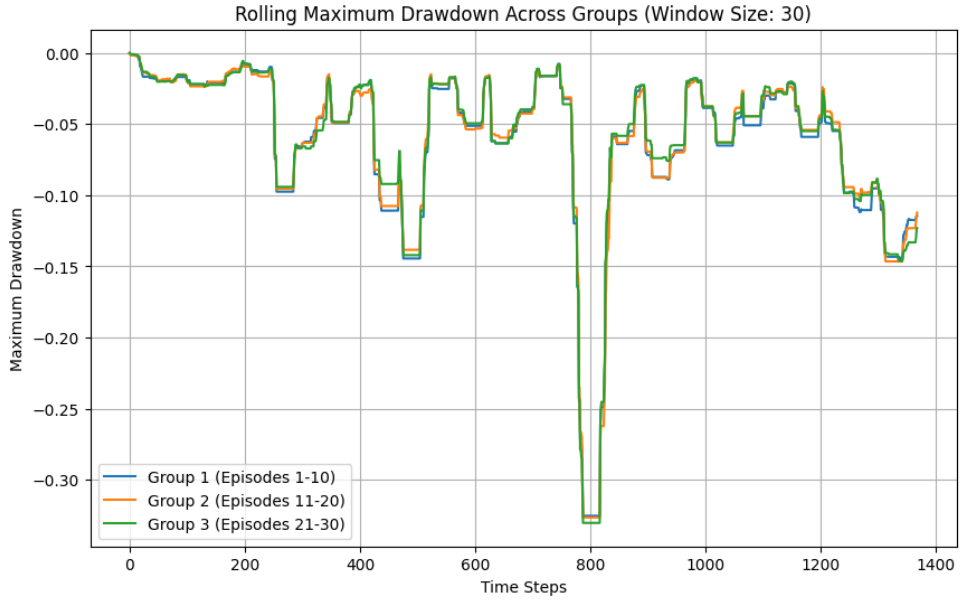

Figure 8: Tau 5 maximum drawdown: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

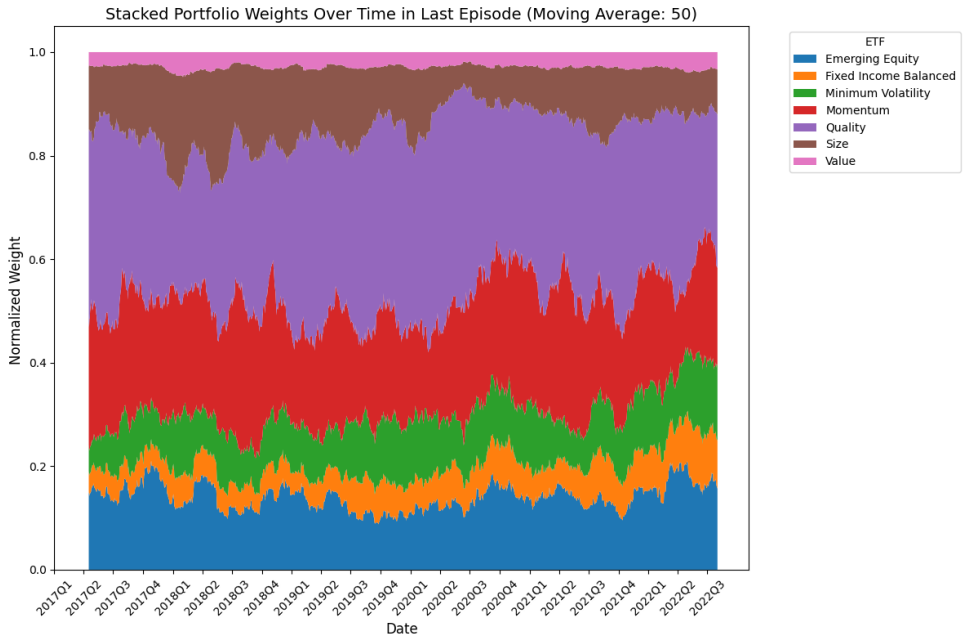

Figure 9: Tau 5 asset allocation: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated

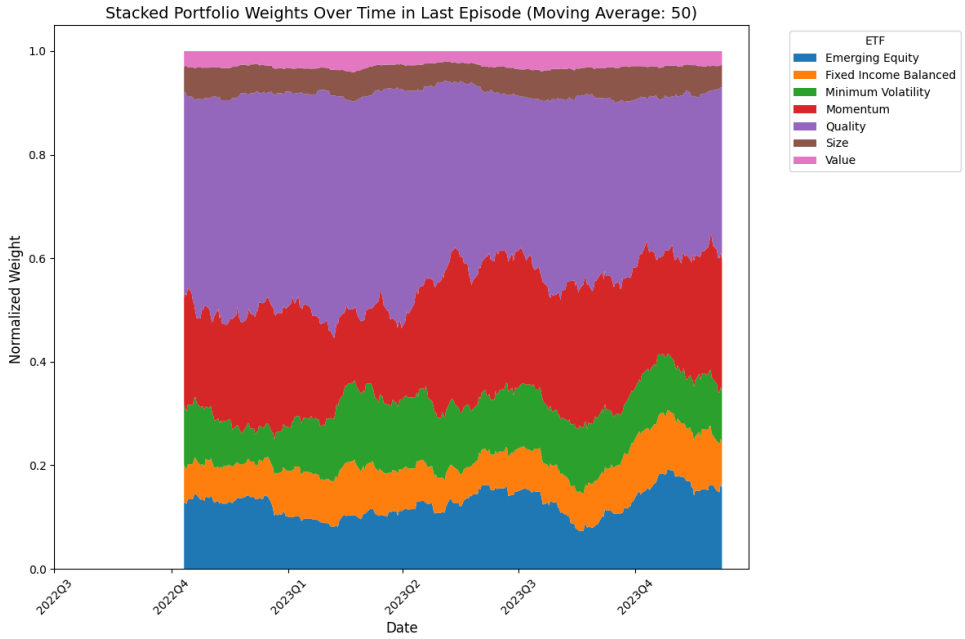

Figure 10: Tau 5 asset allocation: Test set

Note: We used 50-days moving average on portfolio weights while plotting the graph. Both in the train and test set we used 30 episodes and plotted the last episodes outcome.

5.3 Tau 9: Risk-Taker

The allocation results can be seen in Figure 13 and Figure 14. The result is intuitive,the agent neglects low volatility ETFs as they are not providing high returns. The size and momentum effects on risk-return trade-off are well-documented in the Fama-French 5-factor model. Emerging Market ETF has low beta but still has significant volatility, this suggests that it can generate returns even when the overall market is declining. As the agent experienced structural breaks and market corrections, Emerging Market ETFs seems a profitable hedging option. The training results (Figure 24, 27, 25, 26) are similar to the previous

quantiles. When contrasting Tau 9 and Tau 5 losses, we notice a marginal increase in Tau 5 . Beyond the factors discussed earlier (no quantile specific training), it's worth considering the struggle of an economic agent to find the optimal risk-return trade-off. Tau 5, positioned as an intermediate, likely presents more complexities in this aspect compared to the extremes depicted by Tau 2 and Tau 9. This phenomenon likely stems from several factors, including substantial data correlation, the possibility of structural breaks, and the broader discourse surrounding risk diversification across assets. In Figure 15 we plotted the distribution of rewards over the episodes for all the quantiles. As the tau level increases, the distribution of rewards becomes more concentrated around zero. This indicates a stronger emphasis on minimizing extreme outcomes and focusing on risk management.
Over the training process, the agent was slightly improved the accumulated rewards (Figure 11) and compared to Tau 5 the maximum drawdown is higher as it is expected (Figure 12). It seems that agent has more difficulty to find the most profitable strategies as it invested more in risky assets which did not paid off.

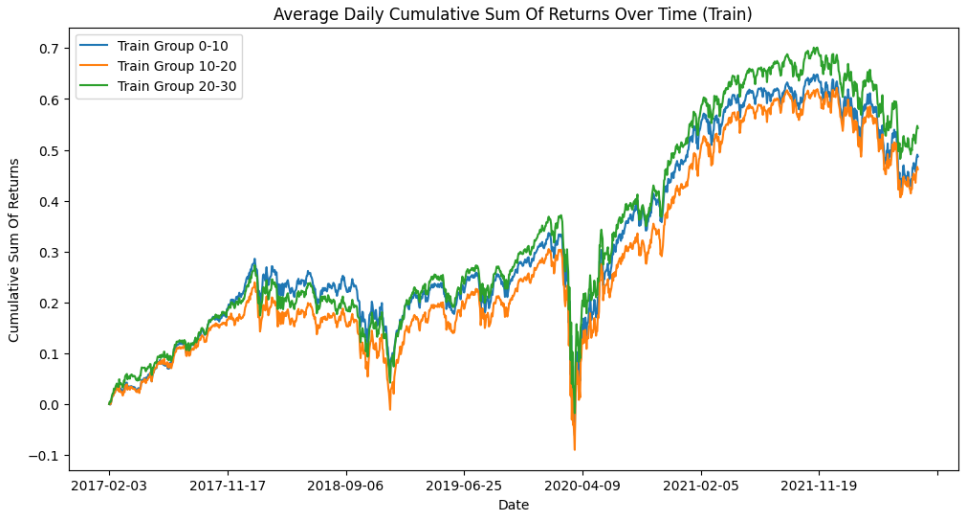

Figure 11: Tau 9 cumulative sum of returns/rewards: Training set
Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

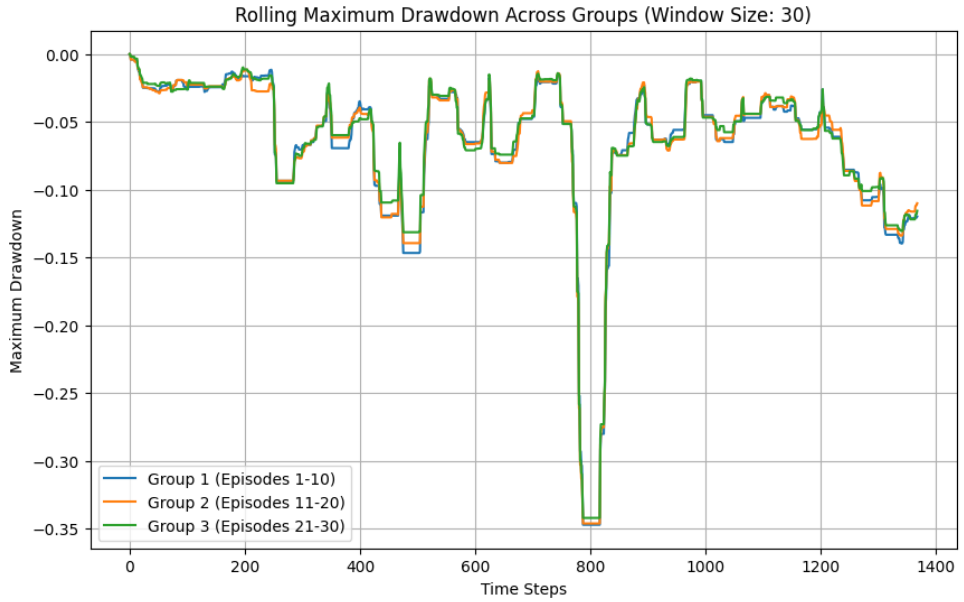

Figure 12: Tau 9 maximum drawdown: Training set
Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

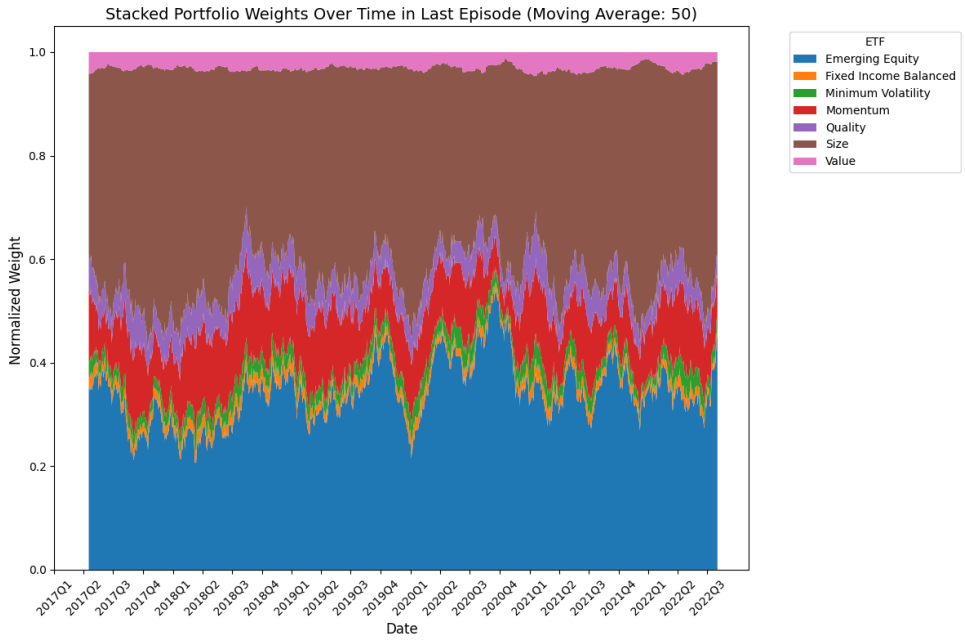

Figure 13: Tau 9 asset allocation: Training set

Note: We used 50-days moving average on portfolio weights while plotting the graph. Both in the train and test set we used 30 episodes and plotted the last episodes outcome.

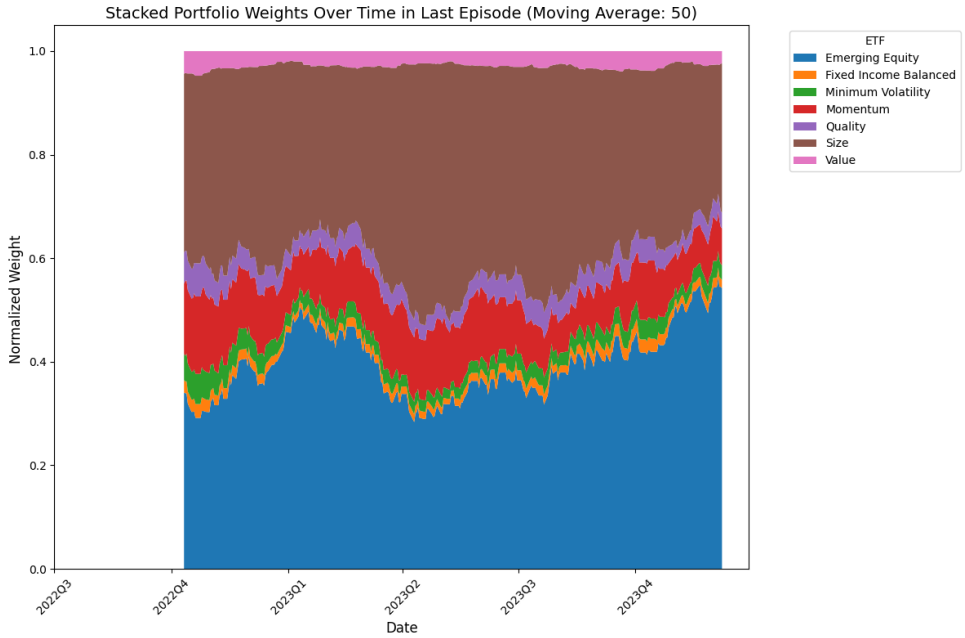

Figure 14: Tau 9 asset allocation: Test set

Note: We used 50-days moving average on portfolio weights while plotting the graph. Both in the train and test set we used 30 episodes and plotted the last episodes outcome.

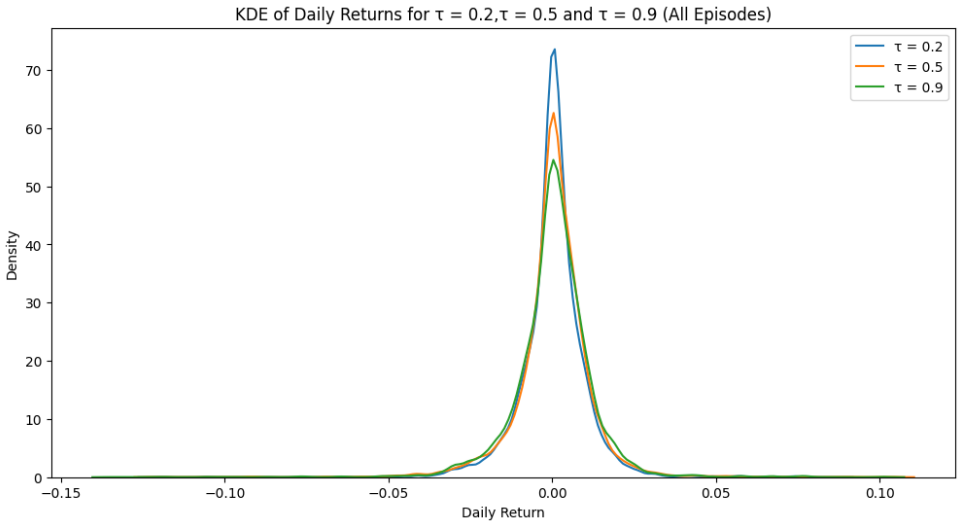

Figure 15: Kernel Density Estimation of Daily Rewards Across Quantiles: Training

6. Conclusion

Previous research in dynamic asset allocation often focused on maximizing expected values, overlooking return distributions or assuming specific distributions for analytical convenience. Asset allocations were commonly equally weighted, lacking economic insights into the training and testing allocation processes. Our contribution fills this gap with a unique on-policy Q-A2C RL model. Unlike DSAC, our model constructs an economically meaningful loss function, facilitating exploration and delivering quantile-specific outcomes. Moreover, we expand the factor investing literature by exploring optimal allocation strategies based on risk preferences.

Lower tau values corresponded to better risk management, evidenced by smaller maximum drawdowns. In Tau 2, asset allocation favored lower beta and standard deviation ETFs, aligning with expectations. However, Tau 5 revealed a different strategy, emphasizing Momentum, Quality, and Size ETFs to optimize risk-return trade-offs, contrary to previous literature suggesting equal-weighted outcomes. Tau 9 incurred higher losses despite allocating more to Size factor ETFs, known for their high risk-premium. While the agents showed learning trends with decreasing loss functions over training, the importance of quantile-specific hyperparameters for capturing risk-profile nuances remained evident.

Our research has many limitations. By construction we did not impose any restriction to the Actor, which implicated high variance of the asset allocation and slower convergence. We did not provide quantile specific models and our on-policy environment can learn sequentially and not from past experience, which makes on-policy models not sample efficient. To extend the research we encourage to use our model with multiple agents to further enhance exploration and faster convergence.

References

André, Eric & Guillaume Coqueret (2020). 'Dirichlet policies for reinforced factor portfolios'. In: *arXiv preprint arXiv:2011.05381*.

Ang, Andrew & Geert Bekaert (2004). 'How regimes affect asset allocation'. In: *Financial Analysts Journal* 60.2, pp. 86–99.

Ba, Jimmy Lei, Jamie Ryan Kiros & Geoffrey E Hinton (2016). 'Layer normalization'. In: *arXiv preprint arXiv:1607.06450*.

Dabney, Will, Mark Rowland, Marc Bellemare & Rémi Munos (2018). 'Distributional reinforcement learning with quantile regression'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Dopfel, Frederick E & Ashley Lester (2018). 'Optimal blending of smart beta and multi-factor portfolios'. In: *The Journal of Portfolio Management* 44.4, pp. 93–105.

Fama, Eugene F & Kenneth R French (2015). 'A five-factor asset pricing model'. In: *Journal of financial economics* 116.1, pp. 1–22.

Feng, Guanhao, Stefano Giglio & Dacheng Xiu (2020). 'Taming the factor zoo: A test of new factors'. In: *The Journal of Finance* 75.3, pp. 1327–1370.

Fons, Elizabeth, Paula Dawson, Jeffrey Yau, Xiao-jun Zeng & John Keane (2021). 'A novel dynamic asset allocation system using Feature Saliency Hidden Markov models for smart beta investing'. In: *Expert Systems with Applications* 163, p. 113720.

Guidolin, Massimo & Allan Timmermann (2007). 'Asset allocation under multivariate regime switching'. In: *Journal of Economic Dynamics and Control* 31.11, pp. 3503–3544.

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel & Sergey Levine (2018). 'Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor'. In: *International conference on machine learning*. PMLR, pp. 1861–1870.

Harvey, Campbell R, Yan Liu & Heqing Zhu (2016). '... and the cross-section of expected returns'. In: *The Review of Financial Studies* 29.1, pp. 5–68.

Hematizadeh, Roksana, Reza Tajaddini & Terrence Hallahan (2022). 'Dynamic asset allocation strategy using a state-dependent Markov model: Applications to international equity markets'. In: *Journal of International Money and Finance* 128, p. 102705.

Khan, Ahmad Zaman & Mukesh Kumar Mehlawat (2022). 'Dynamic portfolio optimization using technical analysis-based clustering'. In: *International Journal of Intelligent Systems* 37.10, pp. 6978–7057.

Khedmati, Majid & Pejman Azin (2020). 'An online portfolio selection algorithm using clustering approaches and considering transaction costs'. In: *Expert Systems with Applications* 159, p. 113546.

Lillicrap, Timothy P, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver & Daan Wierstra (2015). 'Continuous control with deep reinforcement learning'. In: *arXiv preprint arXiv:1509.02971*.

Liu, Xiao-Yang, Zhuoran Xiong, Shan Zhong, Hongyang Yang & Anwar Walid (2018). 'Practical deep reinforcement learning approach for stock trading'. In: *arXiv preprint arXiv:1811.07522*.

Lucarelli, Giorgio & Matteo Borrotti (2020). 'A deep Q-learning portfolio management framework for the cryptocurrency market'. In: *Neural Computing and Applications* 32, pp. 17229–17244.

Ma, Xiaoteng, Li Xia, Zhengyuan Zhou, Jun Yang & Qianchuan Zhao (2020). 'Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning'. In: *arXiv preprint arXiv:2004.14547*.

Markowitz, Harry (1952). 'Portfolio Selection'. In: *The Journal of Finance* 7.1, pp. 77–91. ISSN: 00221082, 15406261. URL: http://www.jstor.org/stable/2975974 (visited on 06/04/2024).

Nazaire, Gregory, Maria Pacurar & Oumar Sy (2021). 'Factor Investing and Risk Management: Is Smart-Beta Diversification Smart?' In: *Finance Research Letters* 41, p. 101854.

Nystrup, Peter, Henrik Madsen & Erik Lindström (2018). 'Dynamic portfolio optimization across hidden market regimes'. In: *Quantitative Finance* 18.1, pp. 83–95.

Staden, Pieter M van, PA Forsyth & Y Li (n.d.). *A data-driven neural network approach to dynamic factor investing*. Tech. rep. Working paper, University of Waterloo, 2021. Available at https://cs ...

Sutton, Richard S & Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, Yichuan Charlie, Jian Zhang & Ruslan Salakhutdinov (2019). 'Worst cases policy gradients'. In: *arXiv preprint arXiv:1911.03618*.

Wang, Mingfu & Hyejin Ku (2022). 'Risk-sensitive policies for portfolio management'. In: *Expert Systems with Applications* 198, p. 116807.

Wang, Yuanrong & Tomaso Aste (2023). 'Dynamic portfolio optimization with inverse covariance clustering'. In: *Expert Systems with Applications* 213, p. 118739.

Yang, Shantian (2023). 'Deep reinforcement learning for portfolio management'. In: *Knowledge-Based Systems* 278, p. 110905.

Ye, Yunan, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, Ju Xiao & Bo Li (2020). 'Reinforcement-learning based portfolio management with augmented asset movement prediction states'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01, pp. 1112–1119.

1. Appendix

Table 2: Definition of Factors

Name	Definition	Characteristics		
		Std (3y)	Equity Beta (3y)	P/B
Value	Includes U.S. stocks with lower valuations based on fundamental characteristics.	18.78 pct	0.93	15.15
Quality	Exposure to U.S. large and mid-cap stocks, emphasizing high return on equity, stable earnings growth, and low debt.	18.93 pct	1.06	28.47
Size	Exposure to U.S. large and mid-cap stocks with relatively smaller average market capitalization.	18.60 pct	1.01	21.54
Momentum	U.S. stocks with higher price momentum in large and mid-cap range.	19.33	0.93	39.22
Minimum Volatility	U.S. stocks with lower risk profiles.	14.20 pct	0.74	22.49
Emerging Market	Stocks in emerging markets with lower volatility characteristics	16.46 pct	0.67	13.73
Fixed Income Balanced Risk Systematic	U.S. bonds, balancing interest rate and credit risk.	7.10 pct	0.30	None

Note: The characteristics column is a snapshot (03/04/2024).The current ETFs' characteristics can be found under 'Portfolio Characteristics' part in each ETFs site.
Source: BlackRock

2. Appendix: Training outcome Tau 2

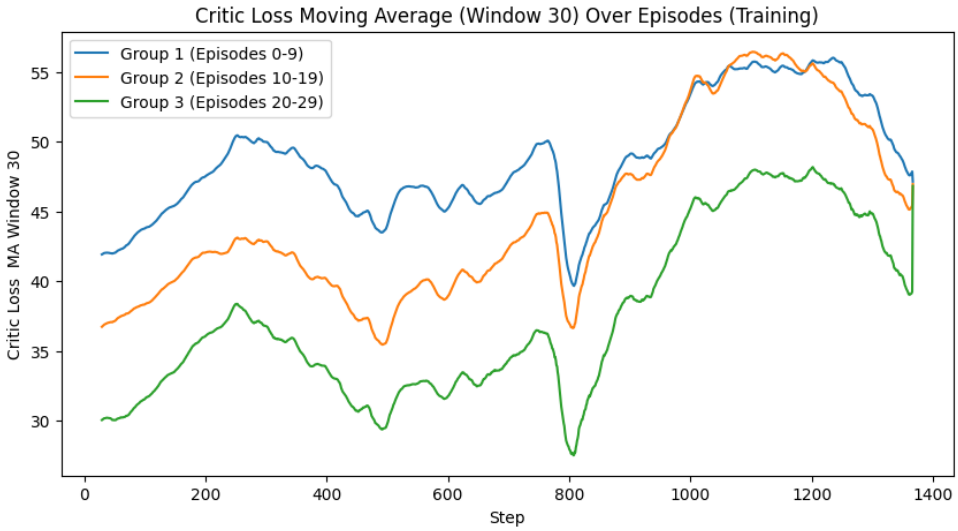

Figure 16: Tau 2 Critic Loss: Training set

Note: The 30 episode outcome is grouped into 3 groups and the group means were calculated.

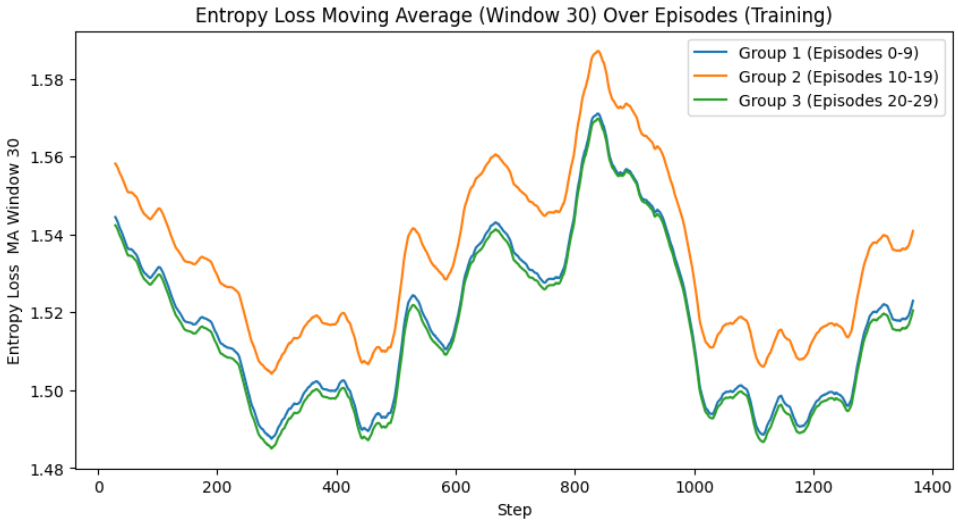

Figure 17: Tau 2 Entropy Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

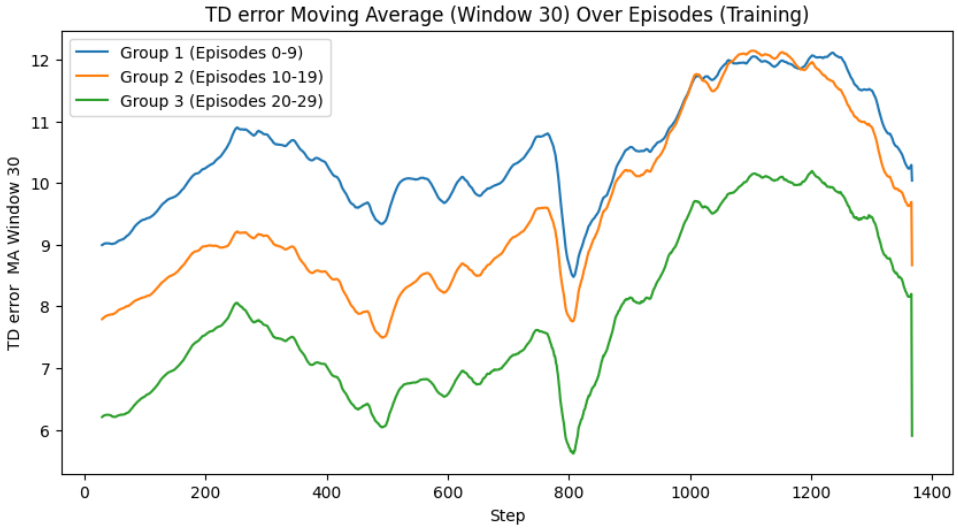

Figure 18: Tau 2 TD Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

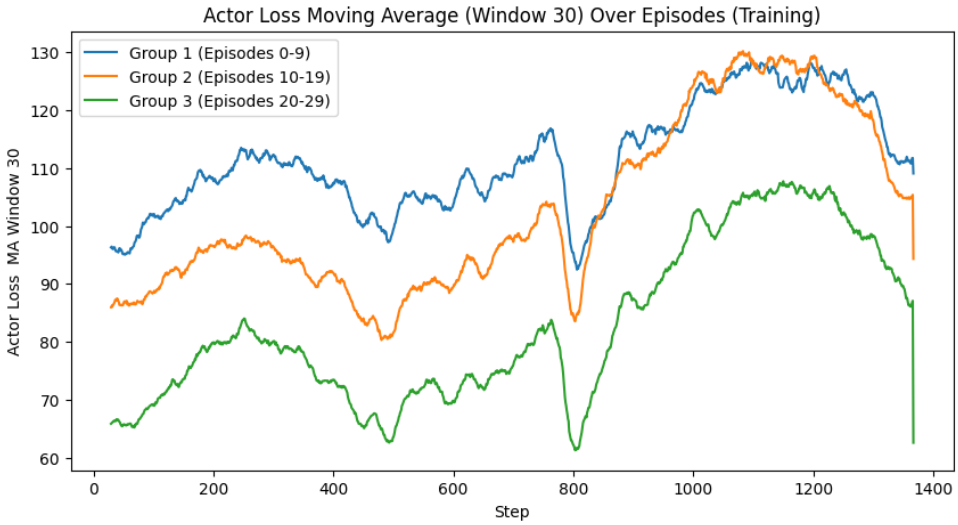

Figure 19: Tau 2 Actor Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

3. Appendix: Training outcome Tau 5

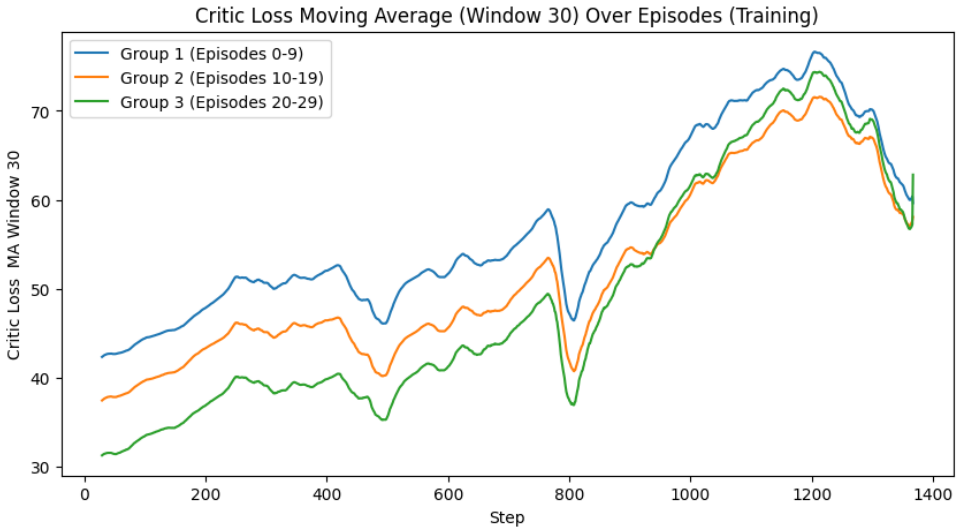

Figure 20: Tau 5 Critic Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

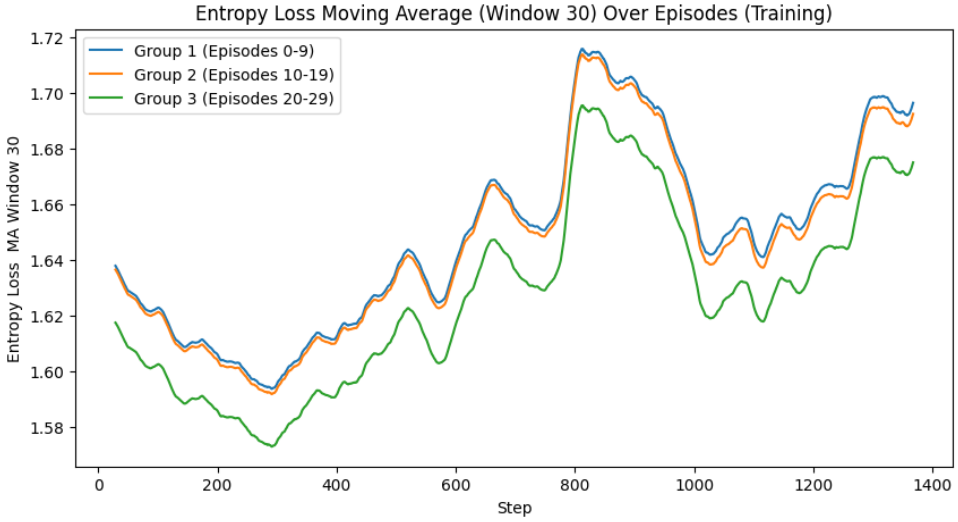

Figure 21: Tau 5 Entropy Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

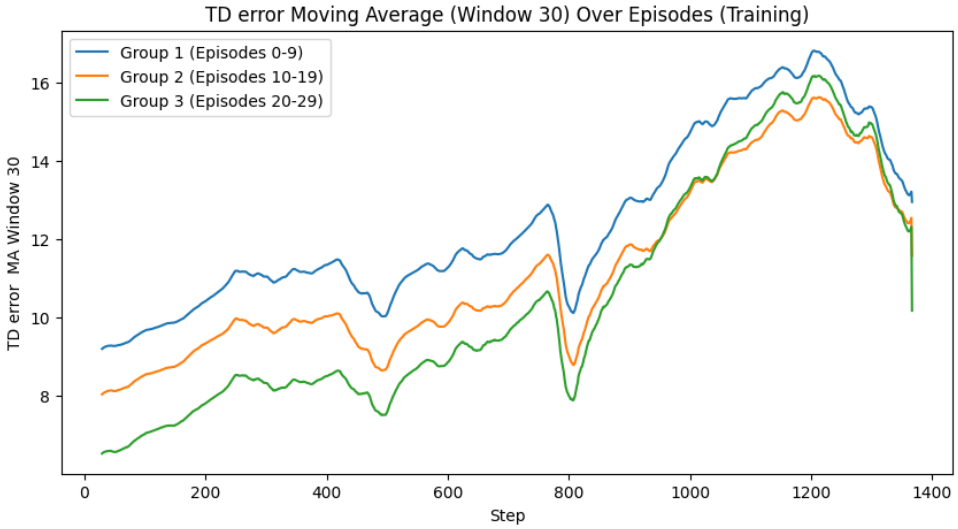

Figure 22: Tau 5 TD Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

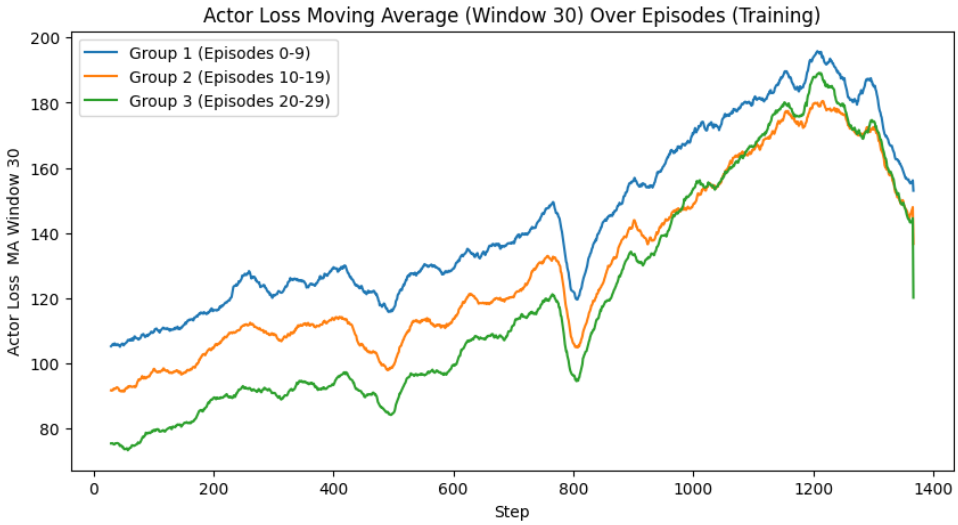

Figure 23: Tau 5 Actor Loss: Training set

Note: The 30 episodes outcome is grouped into 3 groups and the group means were calculated.

4. Appendix: Training outcome Tau 9

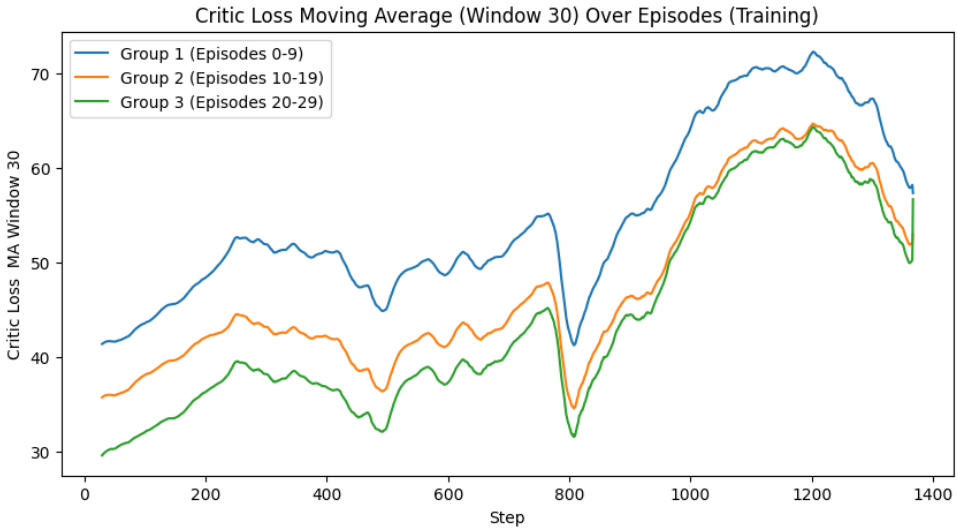

Figure 24: Tau 9 Critic Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

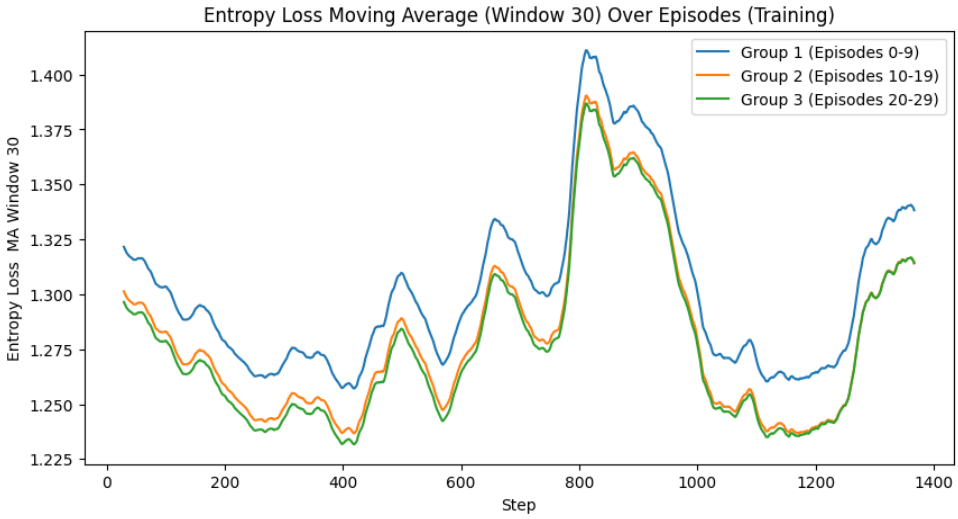

Figure 25: Tau 9 Entropy Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

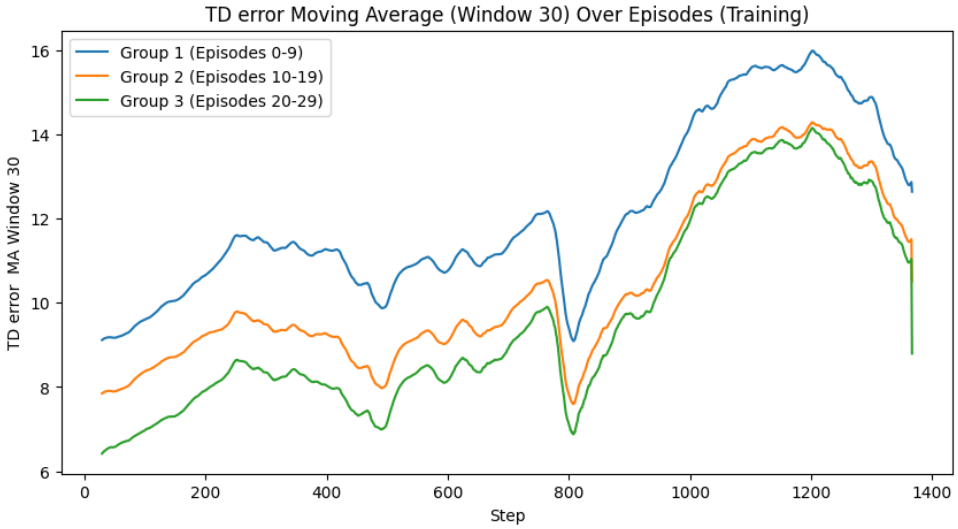

Figure 26: Tau 9 TD Loss: Training set

Note: The 30 episodes' outcome is grouped into 3 groups and the group means were calculated.

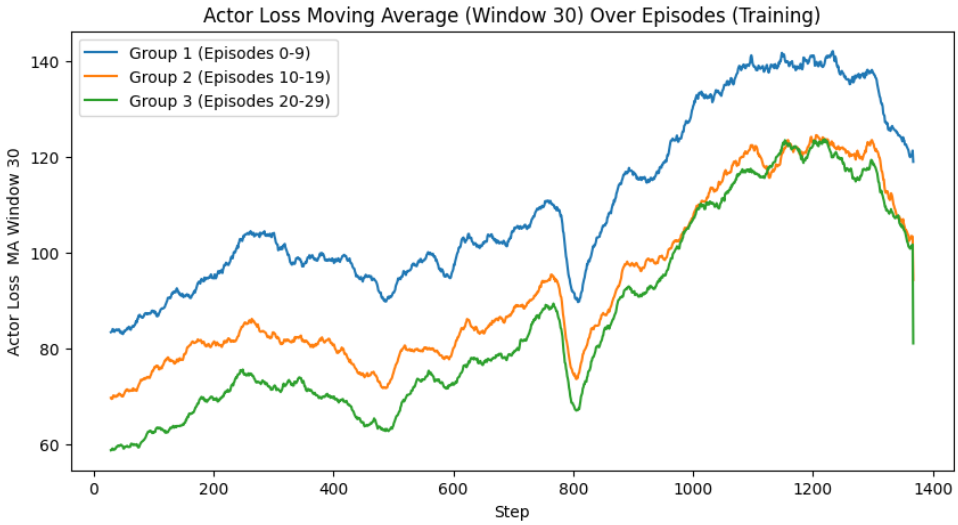

Figure 27: Tau 9 Actor Loss: Training set

Note: The 30 episodes outcome is grouped into 3 groups and the group means were calculated.

IES Working Paper Series

2024

1. Nino Buliskeria, Jaromir Baxa, Tomáš Šestořád: *Uncertain Trends in Economic Policy Uncertainty*
2. Martina Lušková: *The Effect of Face Masks on Covid Transmission: A Meta-Analysis*
3. Jaromir Baxa, Tomáš Šestořád: *How Different are the Alternative Economic Policy Uncertainty Indices? The Case of European Countries.*
4. Sophie Ghvanidze, Soo K. Kang, Milan Ščasný, Jon Henrich Hanf: *Profiling Cannabis Consumption Motivation and Situations as Casual Leisure*
5. Lorena Skufi, Meri Papavangjeli, Adam Gersl: *Migration, Remittances, and Wage-Inflation Spillovers: The Case of Albania*
6. Katarina Gomoryova: *Female Leadership and Financial Performance: A Meta-Analysis*
7. Fisnik Bajrami: *Macroprudential Policies and Dollarisation: Implications for the Financial System and a Cross-Exchange Rate Regime Analysis*
8. Josef Simpart: *Military Expenditure and Economic Growth: A Meta-Analysis*
9. Anna Alberini, Milan Ščasný: *Climate Change, Large Risks, Small Risks, and the Value per Statistical Life*
10. Josef Bajzík: *Does Shareholder Activism Have a Long-Lasting Impact on Company Value? A Meta-Analysis*
11. Martin Gregor, Beatrice Michaeli: *Board Bias, Information, and Investment Efficiency*
12. Martin Gregor, Beatrice Michaeli: *Board Compensation and Investment Efficiency*
13. Lenka Šlegerová: *The Accessibility of Primary Care and Paediatric Hospitalisations for Ambulatory Care Sensitive Conditions in Czechia*
14. Kseniya Bortnikova, Tomas Havranek, Zuzana Irsova: *Beauty and Professional Success: A Meta-Analysis*
15. Fan Yang, Tomas Havranek, Zuzana Irsova, Jiri Novak: *Where Have All the Alphas Gone? A Meta-Analysis of Hedge Fund Performance*
16. Martina Lušková, Kseniya Bortnikova: *Cost-Effectiveness of Women's Vaccination Against HPV: Results for the Czech Republic*
17. Tersoo David Iorngurum: *Interest Rate Pass-Through Asymmetry: A Meta-Analytical Approach*
18. Inaki Veruete Villegas, Milan Ščasný*: Input-Output Modeling Amidst Crisis: Tracing Natural Gas Pathways in the Czech Republic During the War-Induced Energy Turmoil*
19. Theodor Petřík: *Distribution Strategy Planning: A Comprehensive Probabilistic Approach for Unpredictable Environment*
20. Meri Papavangjeli, Adam Geršl*: Monetary Policy, Macro-Financial Vulnerabilities, and Macroeconomic Outcomes*

21. Attila Sarkany, Lukáš Janásek, Jozef Baruník: *Quantile Preferences in Portfolio Choice: A Q-DRL Approach to Dynamic Diversification*

22. Jiri Kukacka, Erik Zila: *Unraveling Timing Uncertainty of Event-driven Connectedness among Oil-Based Energy Commodities*

All papers can be downloaded at: http://ies.fsv.cuni.cz

Univerzita Karlova v Praze, Fakulta sociálních věd
Institut ekonomických studií [UK FSV – IES] Praha 1, Opletalova 26
E-mail : ies@fsv.cuni.cz http://ies.fsv.cuni.cz