INVISIBLE FRONTIERS: ROBUST AND RISK-SENSITIVE FINANCIAL DECISION-MAKING WITHIN HIDDEN REGIMES

MINGFU WANG

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

2023 October

Abstract

In this dissertation, we delve into the exploration of robust and risk-sensitive strategies for financial decision-making within hidden regimes, focusing on the effective portfolio management of financial market risks under uncertain market conditions. The study is structured around three pivotal topics, that is, Risk-sensitive Policies for Portfolio Management, Robust Optimal Life Insurance Purchase and Investment-consumption with Regime-switching Alpha-ambiguity Maxmin Utility, and Robust and Risk-sensitive Markov Decision Process with Hidden Regime Rules. In Risk-sensitive policies for Portfolio Management, we propose two novel Reinforcement Learning (RL) models. Tailored specifically for portfolio management, these models align with investors' risk preference, ensuring the strategies balance between risk and return. In Robust Optimal Life Insurance Purchase and Investment-consumption with Regime-switching Alpha-ambiguity Maxmin Utility, we introduce a pre-commitment strategy that robustly navigates insurance purchasing and investment-consumption decisions. This strategy adeptly accounts for model ambiguity and individual ambiguity aversion within a regime-switching market context. In Robust and Risk-sensitive Markov Decision Process with Hidden Regime Rules, we integrate hidden regimes into Markov Decision Process (MDP) framework, enhancing its capacity to address both market regime shifts and market fluctuations. In addition, we adopt a risk-sensitive objective and construct a risk envelope to portray the worst-case scenario from RL perspective. Overall, this research strives to provide investors with the tools and insights for optimal balance between reward and risk, effective risk management and informed investment choices. The strategies are designed to guide investors in the face of market uncertainties and risk, further underscoring the criticality of robust and risk-sensitive financial decision-making.

Acknowledgements

I would like to express my deepest gratitude to the countless individuals who have played an instrumental role in making this thesis possible. Their unwavering love, support, inspiration, and discipline have shaped me into the person I am today.

First and foremost, I am also profoundly thankful to my supervisors, Hyejin Ku and Yang Shen. Their invaluable guidance, profound expertise, and mentorship have been indispensable. Their wisdom has not only shaped my research trajectory but has also equipped me with the strength to navigate through numerous challenges. I am especially inspired by Hyejin Ku, whose fervor for research and relentless dedication to her students has been a beacon in my academic journey. Her unwavering support and timely encouragement have served as a constant source of motivation, fueling my zeal to learn and grow. I am profoundly indebted to her for the mentorship she provided, which has left an indelible impact on my academic journey. Likewise, I am immensely thankful to Yang Shen for his invaluable counsel and unflinching support. His exceptional expertise and insightful perspectives have been instrumental in refining the contours of my research, guiding me toward achieving my academic objectives. His mentorship has left a lasting impression, for which I am sincerely grateful. The invaluable learnings imparted by both Hyejin and Yang over the years will be a treasure I carry throughout my life. Their lessons extend beyond the confines of the academic sphere and will continue to inspire and guide me in all my future endeavors. This gratitude I feel for them transcends words - I can only hope to honor their trust in me through my achievements.

Furthermore, my sincerest thanks extend to the members of my committee Xin Gao, Steven Wang who have provided me with critical insights and constructive feedback throughout my research process. Their valuable comments and perspectives have added depth and rigor to my work. I am sincerely grateful for their time, effort, and scholarly contributions

that have undoubtedly enriched my thesis.

I am indebted to my parents. Their selfless sacrifices in ensuring my education have been the bedrock of my journey. The ceaseless encouragement and abiding faith they've shown in me have acted as my compass, guiding me toward success. I am truly grateful for their steadfast support and commitment to my growth. In addition, my heartfelt appreciation extends toward my beloved wife Yiwen Liu, whose role in my academic journey cannot be understated. Her ceaseless encouragement and unwavering support throughout this expedition of discovery and learning have been a pillar of strength for me. Her constant companionship has been more than just a source of comfort. It has served as a guiding light, providing clarity and wisdom to both my academic pursuits and personal endeavors. Her resolute faith in me has inspired courage, enabling me to confront challenges head-on, while her ceaseless love has fortified my resolve to continuously strive for excellence.

Lastly, I am deeply appreciative of my dear friends, Chen He, Fan Wang, and my peers, Kaiyan Lin, Richard Le, Yixin Zhang, Hang Du. Their camaraderie, encouragement, and intellectually stimulating discussions have been a constant source of motivation. Their support and friendship have transformed this journey into a memorable and enjoyable experience.

To all those mentioned and to anyone else who has played a part, no matter how big or small, in shaping my academic journey, I offer my sincerest thanks. Without your contributions, this thesis would not have been possible. My heartfelt thanks go to everyone who has accompanied me on this journey. It's been a journey, full of challenges and rewards, that ultimately shaped me into who I am today. Thank you all for believing in me and for being an integral part of my personal growth. This journey would not have been possible without the love, support, and belief of these individuals. I am incredibly fortunate to have had such a supportive and nurturing environment and hope to honor their faith in me by dedicating my achievements to them.

Contents

List of Tables

List of Figures

1 Introduction

Sequential decision-making is a cornerstone in many aspects of financial management and economics. This process of sequential decision-making involves making a series of decisions over a period of time, where each decision can influence subsequent choices and ultimately the final outcome. One of the earliest and most profound applications of sequential decision-making models is in the field of financial management, more specifically, in portfolio management. Portfolio management is an intricate process of choosing and managing an investment policy that minimizes risk and maximizes return on investments. It involves the continuous process of decision-making regarding investment in different assets, balancing the portfolio, and considering the time and uncertainty factors to meet the specific investment objectives. Therefore, the very nature of portfolio management aligns well with the principles of sequential decision-making. Despite its criticality, sequential decision-making is fraught with challenges. The inherent uncertainties of future outcomes are a major concern. Decision-makers must evaluate not only the immediate repercussions of their actions but also anticipate a myriad of potential future scenarios that could be influenced by their choices. The dynamic interplay between immediate decisions and unpredictable future outcomes makes sequential decision-making a captivating and intricate area of study in financial management and economics.

Sequential decision-making models found their initial applications in portfolio management with the advent of the Modern Portfolio Theory (MPT) introduced by Markowitz (1952). MPT proposes that an investment's risk and return characteristics should not be viewed alone, but should be evaluated by how the investment affects the overall portfolio's risk and return. With advancements in technology and the development of sophisticated algorithms, the application of sequential decision-making in portfolio management has become

more robust and powerful. The employment of Reinforcement Learning (RL) in solving portfolio management problems came into light around 1998 when pioneers Moody, Wu, Liao, and Saffell (1998) and Moody and Saffell (2001) are among the first to introduce a recurrent RL algorithm for portfolio optimization problems and develop asset allocation systems, laying the groundwork for future development in portfolio management problems. Following their lead, Almahdi and Yang (2017) extend the recurrent reinforcement learning approach using an adjusted objective function and seek an optimal weight portfolio strategy under the expected maximum drawdown risk measure. In subsequent years, scholars such as Jiang and Liang (2017) and Xiong, Liu, Zhong, Yang, and Walid (2018) present innovative applications of RL technique to tackle the portfolio management problems, focusing specifically on optimal trading positions in the stock and cryptocurrency markets, respectively.

On the other head, the advent of stochastic control in portfolio management opened up an alternative methodology. Stochastic control is a mathematical approach dealing with the optimization of systems that evolve over time under the influence of random disturbances. The stochastic control methodology's intrinsic ability to handle uncertainties and its flexibility in addressing complex systems have rendered it a promising tool in financial portfolio management. Building on these foundational ideas, stochastic control has been used to model and manage dynamic investment strategies, cater to various risk preferences, account for transaction costs, and manage portfolios under various constraints. By providing a systematic and dynamic approach to decision-making, stochastic control allows investors to adjust their strategies based on the current state of the market and anticipated future changes. Merton (1969) is among the first apply the stochastic optimal control theory to develop an elegant solution to the consumption-investment portfolio problem. Richard (1975) extends the models of Merton (1969) by integrating a life insurance purchase decision to the investment-consumption portfolio management problem. This model was tailored for a fixed planning horizon but considered uncertain lifetimes and continuous lifetime income streams. Further refining the use of stochastic control in managing uncertainties, Maenhout (2004) and Elliott and Siu (2009) introduces the concept of robust portfolio rules to handle model uncertainties in the investment world via stochastic control. These robust approaches consider model misspecification in sequential decision-making models, fostering resilience in the

face of uncertainties about the statistical properties of asset returns. These robust method, underpinned by stochastic control, heralds a more adaptive era for portfolio management problem in face of model uncertainties.

Moreover, it's crucial to note that long-term investment strategies are substantially shaped by macroeconomic conditions and business cycles, elements collectively termed as hidden regimes of the market. In addressing these influential factors, Markovian regime-switching models have surfaced as a potent solution. By allowing model parameters to transition between different states over time, these models provide a more realistic representation of financial markets compared to conventional models with fixed parameters. The genesis of these regime-switching models can be traced back to Hamilton (1989), who first develops the model for the stock return time series, called the regime-switching model, which highlights the exceptional capacity of these regimes to capture the complex dynamics of financial markets. Since then, numerous applications of these models have been developed, enhancing the sophistication and financial decision-making by accounting for hidden regimes, and demonstrating that regime-switching models could portray financial markets more accurately than conventional models with deterministic coefficients. Furthering this work, Zariphopoulou (1992) applies the regime-switching model to the investment-consumption problem, in which a financial market consists of a deterministic risk-free asset and a risky asset, depending only on a continuous-time Markov chain. Lee and Shim (2015) apply the Markovian regime-switching model to optimal consumption, investment, and life insurance purchase rules for a wage earner with mortality risk in a continuous time-horizon, and apply the Markov chain approximation method to solve the Hamilton-Jacobi-Bellman (HJB) equation arising from the optimization problem. This thread of research on regime-switching models brings an additional layer of sophistication to financial decision-making, particularly in terms of accounting for hidden regimes. Their ability to switch between different states over time allows these models to capture the impact of structural changes and varying macroeconomic conditions, thereby enhancing the robustness and risk-sensitivity of financial decision-making.

One of the key challenges of portfolio management problems in sequential decision-making models is the existence of hidden regimes. These hidden regimes, often unseen factors or unknown states, can significantly impact the outcomes of our decisions. They might include

underlying shifts in market trends, state transitions, or technological disruptions that are not immediately apparent. Understanding and accounting for these hidden regimes is of utmost importance, as they can profoundly affect the overall performance and success of the strategies we implement. However, integrating hidden regimes in sequential decision-making models is not a straightforward task. It requires not just the ability to detect these regimes, but also to make robust decisions based on the information gleaned from hidden regimes. Markov Decision Processes (MDPs) are widely used framework to model the sequential decision-making problems. Robust MDPs, an extension of MDPs, are designed to handle parameter uncertainties, which are situation some parameters cannot be estimated with precision. The objective of robust MDPs is to seek a policy that maximizes the minimum expected total reward for all possible parameter values, taking into account that the parameter values can fluctuate within an uncertainty set. This ensures robustness against uncertainty and variations in the system. The robust solutions provide a performance guarantee for all uncertain MDP models, thereby offering robustness to model mismatch. Building on this conceptual framework, Iyengar (2005); Nilim and El Ghaoui (2005) formulate robust control of robust MDPs that model an uncertainty set of transition probabilities and derive an optimal policy that performs well under worst-case scenarios using robust dynamic programming. Their contribution informs subsequent research by Wiesemann, Kuhn, and Rustem (2013), who propose a robust MDPs formulation to address the issue of uncertainty in MDPs, where the transition probabilities are unknown or uncertain. They construct a confidence region for the unknown parameters with a specified probability and determine a policy that maximizes the worst-case performance over this region.

Risk-sensitive decisions are another critical aspect of sequential decision-making that are those that explicitly consider and balance the trade-off between the expected outcome and the associated risks. Compared to standard MDPs, risk-sensitive MDPs take into account risk preferences by introducing a risk measure into the objective function, such as variance or Conditional Value-at-Risk (CVaR). This approach allows us to capture the trade-off between the expected return and risk. The goal in a risk-sensitive MDP is to optimize a risk-sensitive objective, such as maximizing the expected return subject to a constraint on the risk. Stella, Lin, and Yan (1998) introduce the risk-sensitive MDPs model, where the objective is to find

a policy that maximizes the probability that the cumulative cost is within some user-defined cost threshold. They propose a Value Iteration (VI) algorithm to solve the problem. However, their algorithm faces scalability issues, limiting its applicability to large-scale problems. Furthermore, Hou, Yeoh, and Varakantham (2014) revisit the risk-sensitive MDPs model and propose a novel approach to solving the problem, called Topological Value Iteration. The new algorithm is more efficient and faster than the original VI algorithm, addressing some scalability concerns. Chow, Tamar, Mannor, and Pavone (2015) consider risk-sensitive MDPs with a CVaR objective, referred to as CVaR MDPs. They provide a new optimization algorithm for CVaR MDPs, which minimize a risk-sensitive CVaR of the total cost in the CVaR MDPs leverages the state augmentation procedure and propose an approximate algorithm with convergence analysis. Lastly, Tamar, Chow, Ghavamzadeh, and Mannor (2016) propose a novel risk-sensitive objective function for RL that considers the consequences of different decisions in a coherent manner. They propose a sampling-based algorithm for estimating the gradient of coherent risk.

In this dissertation, we delve deeply into robust and risk-sensitive approaches to financial decision-making in the realm of portfolio management. Our aim is to construct robust and risk-sensitive policies from the perspectives of reinforcement learning (RL) and stochastic control. We unravel how hidden regimes can influence financial decisions, and explore strategies that navigate these invisible frontiers effectively and robustly. By integrating hidden regimes into MDPs, we endeavor to provide a comprehensive framework that empowers decision-makers to take into account the hidden regimes of the financial market and make more informed decisions in the face of uncertainty. Through this research, we aim to make significant contributions to this expanding body of knowledge, with a particular emphasis on integrating RL and stochastic control with hidden regimes for financial decision-making. We aspire to provide a more nuanced understanding of financial markets, aiding both academia and industry in navigating the complex dynamics of financial decision-making.

The upcoming chapters will delve into the theories and methodologies that underpin our approach, providing support through experimental evidence. In Chapter 2, we design risk-sensitive portfolio management strategies based on the principles of Reinforcement Learning (RL). We introduce two novel approaches for controlling investment risk in portfolio manage-

ment. By leveraging RL techniques, we aim to construct policies that are sensitive to risk, thereby protecting investors from substantial losses. The effectiveness of our approaches is further validated through empirical experiments on real-world data. In Chapter 3, we examine a scenario where an investor is endowed with initial wealth and also receives income continuously over a random lifetime, and she can dynamically purchase life insurance and invest savings in the financial market. This market consists of a risk-free asset and a risky asset, with market coefficients modulated by a continuous-time Markov chain. The states of this chain represent the various regimes of the financial market. In this context, we seek solutions for robust optimal life insurance and investment-consumption strategies under the framework of regime-switching. In Chapter 4, we explore the integration of hidden regimes into the Markov Decision Process (MDP). We propose a novel approach to address the robust and risk-sensitive MDP problem with hidden regimes from both RL and Stochastic Control (Dynamic Programming) perspectives. We use the current state and the financial market's hidden regimes to model uncertainty over transition probabilities, thereby constructing robust and risk-sensitive policies.

In traversing this intricate landscape, we aim to illuminate ways to better manage risk and make decisions robust against unseen and unknown factors. It is our hope that this contributes to the broader discourse on financial decision-making in our modern era, providing tangible strategies to navigate the complex financial markets.

2 Risk-sensitive Policies for Portfolio Management

In this chapter, building upon our findings as published in M. Wang and Ku (2022), we delve deeper into the realm of risk-sensitive policies for portfolio management. We aim to unpack the intricacies of these policies, exploring the performance of RL algorithms in portfolio management problem.

2.1 Introduction

Portfolio management is a decision-making process that allocates investment funds to gain maximum profit and relatively lower risk based on individuals' goals, risk preferences, and investment horizons. The foundation of modern portfolio theory can be traced back to the pioneering work of Markowitz (1952), in which his Mean-Variance analysis is a representative methodology in the framework of return-risk trade-off analysis. The original Mean-Variance theory is developed in a mathematical skeleton that constructs a portfolio that maximizes the expected return for a given degree of risk. The disadvantage of the original mean-variance model is that when a portfolio has high return and volatility, investors might give up the strategy of high returns to remain at low risk. Moreover, there is much noise and uncertainty in the financial market, which leads to inaccurate values of the mean and variance.

Over the last few decades, many studies investigate the application of RL algorithms to financial market trading, and try to predict the price movements or trends by using historical market data. The benefit of RL learning technology in portfolio optimization problems is that the algorithm can observe and learn from the market history completely without assuming any prior knowledge of the financial markets or making any models. The advantage of RL learning technology in portfolio optimization is that the algorithm can thoroughly learn and extract useful information from market history, without any advanced knowledge and experience in

financial markets, and without making any assumptions about the models. The application of RL in portfolio management problems starts from 1998, Moody et al. (1998) and Moody and Saffell (2001) first propose a recurrent RL algorithm for portfolio optimization problem and construct assets allocation systems. Both of these studies aim to maximize the differential Sharpe ratio, that is, to maximize risk-adjusted returns by considering transaction costs. The disadvantage of using the differential Sharp ratios is that it penalizes returns exceeding a certain value and takes more weight on recent returns. In addition, the differential Sharpe ratio cannot distinguish the potential growth trend of the portfolio. In another attempt by Almahdi and Yang (2017), they extend the recurrent reinforcement learning approach using an adjusted objective function and seek an optimal weight portfolio strategy under the expected maximum drawdown risk measure. However, these existing models have fixed the number of shares for trading. In reality, when the buying or selling signal occurs in the market, it is necessary to determine how many shares to buy or sell. Trading a fixed number of stocks in each transaction does not reflect the real market situation and affects the total profits. With the development of deep RL, deep RL has demonstrated the capability to learn complex policies from many types of environments.

Deep Q-network (DQN) is one of the most popular methods in deep RL. As the approximation of the Q-value function, the neural network can be applied to approximate the reward by taking actions and pursuing policies from a given state. Bertoluzzo and Corazza (2012), Chen and Gao (2019), and Park, Sim, and Choi (2020) apply DQN to portfolio management problems and make remarkable achievements. The advantage of DQN is that it does not require the labelled data that suffer from the constraints and bias of data. It can automatically adapt to the changes in the underlying data distribution, thereby it is a suitable method for the dynamic of the financial market. However, their actions are limited to the discrete action space while the actions are continuous in portfolio management problems. To overcome this issue, DDPG is proposed by Google DeepMind (Lillicrap et al., 2015), a type of actor-critic based DRL algorithm that supports the continuous action space encountered in portfolio optimization problems. Jiang and Liang (2017) and Xiong et al. (2018) present innovative approaches based on DDPG to solve the trading problem of the optimal trading position at each transaction in stock market and cryptocurrency market. The experimental results of

their evaluation take into consideration transaction costs and prove the effectiveness of the algorithms in portfolio management. The advantage of DDPG is that it can deal with the problem of high-dimensional continuous action space well, and its purpose is to learn a policy function directly, instead of approaching the Q-value function.

Hierarchical Learning (HRL) is a promising method that expands the traditional reinforcement learning methods by decomposing the elaborate and intricate problems into sub-problems and effectively solving each sub-problem. The HRL method has some advantages, such as it is easier to be trained, and solving each sub-problem individually will improve its reusability, which will accelerate the learning process. HRL has been devoted to learning these difficult tasks for a long time, the multi-layer strategies are trained to make decisions and control at a higher level of temporal and behavior abstraction. (Barto & Mahadevan, 2003; Dayan, 2002; Dietterich, 1998; Nachum, Gu, Lee, & Levine, 2018). In general, by having a hierarchy of policies, only the lower-level policies execute actions to the environment, and the higher-level policies are trained to distribute the sub-tasks to the lower-level policies. Although the applications of deep RL algorithms in the financial market are well studied, most of the previous works only consider maximizing the total profit, and surprisingly, they ignore the impact of possible disasters.

Managing risk in dynamic decision-making is an important topic because it can fully identify and deal with potential risks. A well-known approach is to consider risk while measuring the performance of a trading strategy. The risk is a quantity related to the variance (or standard deviation) of the rate of return, and it is also referred to as volatility. The Sharp ratio is one of the most popular indicators that consider the profits generated by trading strategies and the risk associated with trading strategies. There are existing works that proposed risk-sensitive portfolio optimization algorithms. Among them, Schlosser (2020) proposes risk-sensitive trading strategies based on intraday trading, it allows to control the moments of the reward distribution. Then, the author uses dynamic programming techniques to express recursively the higher moments of reward distribution and obtain the optimal solutions. Y. Gao, Lui, and Hernandez-Leal (2021) present the Risk-Averse Averaged Q-Learning and Variance Reduced Risk-Averse Q-Learning by combining the risk-averse functions and variance reduction techniques. Then, they augment the framework to a multi-agent scenario and

propose Risk-Averse Multi-Agent Q-Learning (RAM-Q) for trading markets that augment multi-agent with robustness. Harnpadungkij, Chaisangmongkon, and Phunchongharn (2019) propose a risk-sensitive algorithm by applying distributional reinforcement learning, called C21-SR, which models cumulative returns using a 21-bin discrete distribution and selects the actions according to the Sharpe ratio to control investment risk and maximize profits.

In this chapter, we aim to construct policies with risk awareness to protect the investor under the worst-case scenarios. Inspired by Dietterich (1998); Nachum et al. (2018), we propose a novel RL algorithm, called Hierarchical DDPG, which combines the classical DDPG algorithm and the Hierarchical structure for portfolio management problems. The original higher-level policy of HRL performs at an abstraction layer and distributes sub-tasks to the lower-level policy, which correspond directly to the target that the lower-level policy attempts to reach. In our proposed Hierarchical DDPG, the higher-level policy adjusts the lower-level policy's actions to reduce the portfolio risk and operates in the environment. We employ parametric Conditional Value-at-Risk (CVaR) as a metric that measures the portfolio risk. HRL is extended by adding the portfolio risk indicator, so that the agent can implement different trading strategies for different scenarios. More precisely, the lower-level policy of Hierarchical DDPG can be interpreted as a *worker*, aiming to maximize the total profit of the portfolio when the portfolio risk is lower than the CVaR constraints. The higher-level policy of Hierarchical DDPG can be interpreted as a *manager*, whose purpose is to reduce the portfolio risk immediately based on the *worker*'s action when portfolio risk exceeds the investor's tolerance. On the other hand, most of the existing RL algorithms cannot learn risk-sensitive policies because they only consider maximizing the average and do not penalize the effects of rare occurrences of catastrophic events. Motivated by Barth-Maron et al. (2018); Tang, Zhang, and Salakhutdinov (2020), we propose the distributional DDPG model for portfolio management problems with the purpose of seeking a risk-sensitive policy that can map the same state to different actions according to risk preference. We construct the α-*percentile* expectation as our measure, which represents the expected return under the distribution of the α-*percentile* at the bottom of future return. The risk-sensitive policies can be obtained by maximizing the α-*percentile* expectation based on different values of risk parameter α. When α is small, the agent focuses on maximizing the performance of the

worst-case scenario.

The main goal of this chapter is to construct a risk-sensitive policy to protect investors who may suffer a huge loss due to a financial crisis or rare disaster events. In pursuing this goal, we have made the following contributions. First, we design the Hierarchical DDPG algorithm to learn the solution of the portfolio management problem. When the portfolio risk is below the CVaR constraints, the Hierarchical DDPG agent aims to maximize the total profit. But when the portfolio risk exceeds the CVaR constraints, the priority of the Hierarchical DDPG agent is to reduce the portfolio risk immediately, instead of maximizing the total profit. Second, we propose the distributional DDPG method for solving the portfolio optimization problem based on uncertainty of the future returns. According to the investor's risk preference, the distributional DDPG algorithm can learn a risk-averse policy that yields different actions depending on risk parameters, which is more robust than the other RL algorithms.

The proposed approaches are then validated by a real-world dataset from the U.S. stock market[1]. It is well known that the U.S. stock market crashed during the Coronavirus pandemic in 2020, which was one of the most dramatic stock market crashes in history. The circuit breaker mechanism was triggered three times in a month, S&P500 plunged 1019 points, an equivalent of roughly 29%. This provides a good example for verifying our algorithms. The three different comprehensive performance metrics are employed to assess the portfolio performance from different perspectives. Our experimental results show that Hierarchical DDPG is superior in portfolio management to the classical DDPG method because it can significantly reduce or avoid a loss caused by the occurrences of catastrophic events. Also, the results demonstrate that the distributional DDPG agent can provide a risk-averse policy depending on the risk parameter, and the α-*percentile* expectation is well-suited as the criterion of the distributional DDPG, which provides a good distributional critic that can be learned. Via the experimental study, we verify that two proposed algorithms provide promising results, and our approaches are an effective way to protect the investor who may suffer a huge loss due to a financial crisis or rare disaster events.

This chapter is organized as follows. Section 2.2 briefly reviews the related work in the area

[1]Real data is collected from Yahoo Finance.

of portfolio management using RL. Section 2.3 formulates the portfolio allocation problem. Section 2.4 introduces the classical DDPG algorithm. Section 2.5 introduces our proposed novel models. The core innovation of this chapter, Hierarchical DDPG and Distributional DDPG for the portfolio management problem is presented in this section. Section 2.6 displays the experiment results for classical DDPG, Hierarchical DDPG, and Distributional DDPG; and analyzes the obtained results. The final conclusion is presented in section 2.7.

2.2 Related Work

In recent decades, portfolio optimization problems in financial trading have attracted much attention. As a primary approach in the field of Artificial Intelligence (AI)[2], deep RL is one of the most popular portfolio management methods in the financial market due to its outstanding performance compared to expert traders. Deep RL has originally been used in applications of video games (Mnih et al., 2015) and chess games (Silver et al., 2018). Ormoneit and Glynn (2002) propose a kernel-based RL method to conquer the issue of instability in RL. Their method aims at learning within the framework of average-cost and applying this method to portfolio management problems. Nevmyvaka, Feng, and Kearns (2006) propose a novel RL algorithm for optimizing transaction execution in the modern financial market by using NASDAQ market high-frequency data sets. Most traders in the real world are dealing with large-scale diversified investment portfolios, but due to time constraints, they are always neglected to deal with individual stock and millisecond data, which makes it necessary to use automatic trading agents. Their experiment results of real-world data on three NASDAQ stocks demonstrate that RL can indeed result in significant improvements.

Deep RL that combines deep learning and RL algorithms can divide into three groups: policy gradient, value-based, and actor-critic. Policy gradient algorithms learn directly the stochastic policy function that maps a state to the probability of each action in action space. Value-based algorithms approximate the Q-value function that represents the expected accumulated rewards by given a state on taking an action and pursuing a policy. The observed

[2]In its 12th annual Global Alternative Fund and Investor Survey, November 2018, Ernst & Young (EY) reports that more than 40% of hedge fund managers admit that they refer to AI to develop strategies to enhance performance for making greater profits in their investment process.

information is analyzed through the neural network and output Q-values of each action, then the value-based algorithms rely on the reward function to influence the output of neural networks by backpropagation. The Temporal difference (TD) learning method plays a key role in the actor-critic algorithm that combines the value-based method and the policy-based method. The policy-based network plays as an actor who outputs an action, while the value-based network acts as a critic that appraises the action estimated by the actor-network and generates the TD errors to update the actor and critic network.

DQN is one of the value-based deep learning methods, which updates the Q-value through a neural network instead of updating the Q-table to maximize the cumulative rewards. In the absence of a deterministic strategy, the algorithm will select the action that provides the highest Q-value, and then the Q-value will be updated continuously until it converges to the best action. Bertoluzzo and Corazza (2012) apply DQN to portfolio management problems. The action space is defined as buying, selling or hold. To compare the performance of DQN and Kernel-based RL algorithm, real-world data from three Italian stocks are used to test and validate the performance. The experiment results show the DQN algorithm performs better than the Kernel-based RL algorithm. Chen and Gao (2019) combine the DQN and Deep Recurrent Q-network for portfolio optimization problems and construct a daily stock trading system that can automatically decide to make transactions on each trading day. The Standard & Poor's 500 Index ETF is used to evaluate their trading system, and its daily prices are defined as the state of reinforcement learning in the trading environment. Jeong and Kim (2019) propose an automated system that can predict the number of shares of each transaction by adding a DNN regressor to DQN. In addition, they adopt a transfer learning technique to pre-train neural networks when financial data is insufficiently large. The experiment results reveal that the total profit is significantly increased by forecasting the number of shares. Pendharkar and Cusatis (2018) design an on-policy SARSA and off-policy Q-learning for the purpose of asset allocation, train the RL agent with discrete action space, which can maximize the return of portfolio or differential Sharp ratio, and compare it with other RL methods in financial markets. Z. Gao, Gao, Hu, Jiang, and Su (2020) propose a novel DQN framework, which is expressly designed for managing a multi-asset portfolio and allows DQN agents to optimize their trading strategies by interacting with the real

financial market. Park et al. (2020) derive a novel portfolio trading strategy for multi-asset management in the practical action space and devise a transformation function that maps the infeasible action to similar feasible actions.

Deep Deterministic Policy Gradient (DDPG), proposed by Lillicrap et al. (2015), is one of the actor-critic algorithms that support continuous action space. Compared to DQN, the merit of DDPG is that it can handle high-dimensional continuous action problems well, and it directly outputs the optimal action instead of the Q-value. Jiang and Liang (2017) implement the DDPG algorithm that adopts a convolutional neural network (CNN) to solve the asset allocation problem in the cryptocurrency market. They optimize the investment portfolio by weighting all stocks, and make it suitable for continuous-time actions to solve the discrete action space problem. A back-test experiment is applied in the cryptocurrency market, and their experimental results achieve positive results compared to another three RL portfolio management algorithms. Liang, Chen, Zhu, Jiang, and Li (2018) extend DDPG by using a deep residual network and propose an adversarial training method that improves the performance of deep RL. It has been tested on the Chinese stock market that illustrates this approach can significantly improve the training efficiency, average daily earnings and Sharp ratio. Xiong et al. (2018) explore the potential of training DDPG agents to obtain the optimal trading strategy in the stock market. They construct a portfolio that consists of 30 stocks, and the trading environment has been created by adopting the daily prices of each stock. Compared to the traditional minimum-variance method, the DDPG algorithm has gained higher benefits, which proves the effectiveness of the algorithm. Wu and Li (2020) construct Gate Deterministic Policy Gradient (GDPG) by adding the Gate Recurrent Unit into DDPG to extract financial features from the time-series stock market data. The performance of their proposed GDGP method is verified by comparing the experimental results, which shows that the GDPG method gains a higher return than the traditional DRL, it can spawn a more stable performance even in the turbulent financial market.

Despite a tremendous amount of research and high-quality results in the area of portfolio management by RL, there is very little literature that takes into account the portfolio risk, especially under the worst-case scenarios, in constructing trading strategies. Every financial product has its own risk and reward characteristics. The ultimate goal of investors is to

choose the best portfolio with the highest return, and keep the portfolio risk below a certain degree. Thus, it is significant and important for the RL agent to construct a risk-sensitive or risk-averse policy for the investors who may suffer a huge loss caused by rare events.

2.3 Portfolio Allocation Problem

Portfolio optimization requires continuous reallocation of an investment fund into different assets. Our trading agent does this allocation periodically. The trading environment is formulated as follows. For the convenience of readers, we provide Table 2.1 that includes all symbols.

2.3.1 Problem formulation

The individual asset consists of the opening, highest, lowest, closing prices, and volume for each trading period. We denote the closing price $v_{i,t}^c$ of the i-th asset in the t-th trading period. Similarly, $v_{i,t}^h$, $v_{i,t}^l$, $v_{i,t}^o$ denote the highest, lowest, and opening prices of the i-th asset in the t-th period, respectively. Denote by $v_{i,t}^v$ the volume of the i-th asset in the t-th period. For the t-th trading period, the prices and volume of each individual asset can be expressed as

$$v_{i,t} = \left[v_{i,t}^o, v_{i,t}^h, v_{i,t}^l, v_{i,t}^c, v_{i,t}^v\right], \tag{2.1}$$

and the information the agent can observe on the i-th stock at timestep t is written as

$$V_{i,t} = \begin{bmatrix} v_{i,t}^o, & v_{i,t}^h, & v_{i,t}^l, & v_{i,t}^c, & v_{i,t}^v \\ v_{i,t-1}^o, & v_{i,t-1}^h, & v_{i,t-1}^l, & v_{i,t-1}^c, & v_{i,t-1}^v \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ v_{i,t-m+1}^o, & v_{i,t-m+1}^h, & v_{i,t-m+1}^l, & v_{i,t-m+1}^c, & v_{i,t-m+1}^v \end{bmatrix}, \tag{2.2}$$

where m is the window size.

For continuous markets, the relative price change in the t-th trading period is defined as the element wise division of $v_{i,t}^c$ by $v_{i,t-1}^c$:

$$y_t = \frac{v_t^c}{v_{t-1}^c} = \left[1, \frac{v_{1,t}^c}{v_{1,t-1}^c}, \frac{v_{2,t}^c}{v_{2,t-1}^c}, \ldots, \frac{v_{n,t}^c}{v_{n,t-1}^c}\right]^T, \tag{2.3}$$

Table 2.1: Notations for the trading system

Symbols	**Explanations for the notation**
$v_{i,t}^c$	closing price of the i-th asset in the t-th trading period
$v_{i,t}^h$	highest price of the i-th asset in the t-th trading period
$v_{i,t}^l$	lowest price of the i-th asset in the t-th trading period
$v_{i,t}^o$	opening price of the i-th asset in the t-th trading period
$v_{i,t}^v$	volume of the i-th asset in the t-th trading period
$v_{i,t}$	prices and volume of the i-th asset in the t-th trading period
w_t	portfolio weight at the beginning of the period $t+1$
w_t'	portfolio weight at the end of the period t before the execution action
p_t	portfolio value at the beginning of the period $t+1$
p_t'	portfolio value at the end of the period t before the execution action
y_t	relative price change in the t-th trading period
r_t	rate of return at the end of the t-th trading period
$\hat{\mu}_t$	expected value of the return at the end of the t-th period
μ_t	mean value of portfolio return at the end of the t-th period
$\hat{\Sigma}_t$	variance-covariance matrix of return at the end of the t-th period
$\mathscr{V}_t$	variance of portfolio return at the end of the t-th period
c	commission rate for buying and selling
C_t	trading cost rate for the t-th trading period
n	number of risky assets
m	window size

where n is the number of stocks in the portfolio. Note that the first element of y_t represents the relative price of cash, therefore, it is always 1. We can use this relative price change vector to calculate the portfolio value in a period. The portfolio weight vector is defined as

$$w_t = [w_{0,t}, w_{1,t}, w_{2,t}, \ldots, w_{n,t}], \tag{2.4}$$

where $w_{i,t}$ is the fraction of investment on stock i at the beginning of period $t+1$ with the initial portfolio weight vector $w_0 = [1, 0, 0, ..., 0]$, and $\sum_{i=0}^{n} w_{i,t} = 1$ with each $w_{i,t} \geq 0$. Note that the initial value of the weight vector w_0 indicates that all the investment capital is in the riskless asset at the beginning. Assuming p_{t-1} is the portfolio value at the beginning of period t, ignoring transaction costs, the portfolio value at the end of period t can be calculated as

$$p_t = p_{t-1} y_t \cdot w_{t-1}, \tag{2.5}$$

where w_{t-1} is the portfolio weight vector at the beginning of period t and its i-th element $w_{i,t-1}$ is the proportion of stock i in the portfolio after capital reallocation.
The rate of return at the end of period t can be calculated as

$$r_t = \frac{p_t}{p_{t-1}} - 1 = y_t \cdot w_{t-1} - 1, \tag{2.6}$$

and the corresponding logarithmic rate of return is given by

$$\log(r_t) = \log(\frac{p_t}{p_{t-1}} - 1) = \log(y_t \cdot w_{t-1} - 1). \tag{2.7}$$

The mean value and variance of portfolio return can be formulated as

$$\mu_t = w_t \cdot \hat{\mu}_t, \tag{2.8}$$

$$\mathscr{V}_t = w_t^T \hat{\Sigma}_t w_t, \tag{2.9}$$

where $\hat{\mu}_t$ and $\hat{\Sigma}_t$ are the mean value and the variance-covariance matrix of return for each asset. Note that the $\hat{\mu}_t$ and $\hat{\Sigma}_t$ are estimated every step based on the observation and window size. If there is no transaction cost, the final portfolio value will evolve as follows

$$p_T = p_0 \prod_{i=1}^{T} (1 + r_i) = p_0 \prod_{t=1}^{T} y_t \cdot w_{t-1}, \tag{2.10}$$

where p_0 is the initial investment amount.

2.3.2 Transaction costs

In the real world, buying or selling assets incurs a transaction cost, usually in the form of commission fee. Assuming a constant commission rate, we can recalculate the final portfolio value. At the beginning of period t, the portfolio's action vector is w_{t-1}. Due to the price changes of assets in the market, the portfolio weight vector transforms to w'_t at the end of period t:

$$w'_t = \frac{y_t \odot w_{t-1}}{y_t \cdot w_{t-1}}, \tag{2.11}$$

where $\odot$ is element-wise multiplication. The mission of the agent is to reallocate portfolio weights from w'_t to w_t by buying or selling relevant assets. Paying all commission fees, this reallocation action shrinks the portfolio value. If we set a constant commission rate $c \in [0, 1)$ for buying and selling, then the trading cost rate of each period C_t can be approximated as (Hegde, Kumar, & Singh, 2018; Jiang & Liang, 2017):

$$C_t = c \sum_{i=1}^{n} \left| w'_{i,t} - w_{i,t} \right|, \tag{2.12}$$

where $C_t \in [0, 1)$. Assuming all buying and selling trades are executed at the end of day, the portfolio value (2.5) at the end of day t evolves

$$p_t = (1 - C_t) p'_t, \tag{2.13}$$

where p'_t represents the portfolio value at the end of period t before execution, that is, $p'_t = p_{t-1} y_t \cdot w_{t-1}$.

Therefore, the rate of return (2.6) can be rewritten as:

$$r_t = \frac{p_t}{p_{t-1}} - 1 = (1 - C_t) y_t \cdot w_{t-1} - 1. \tag{2.14}$$

Hence, the final portfolio value can be expressed as

$$p_T = p_0 \prod_{t=1}^{T} (1 - C_t) y_t \cdot w_{t-1}. \tag{2.15}$$

Figure 2.1 demonstrates the dynamic relationships among portfolio values and weight vectors on the time axis.

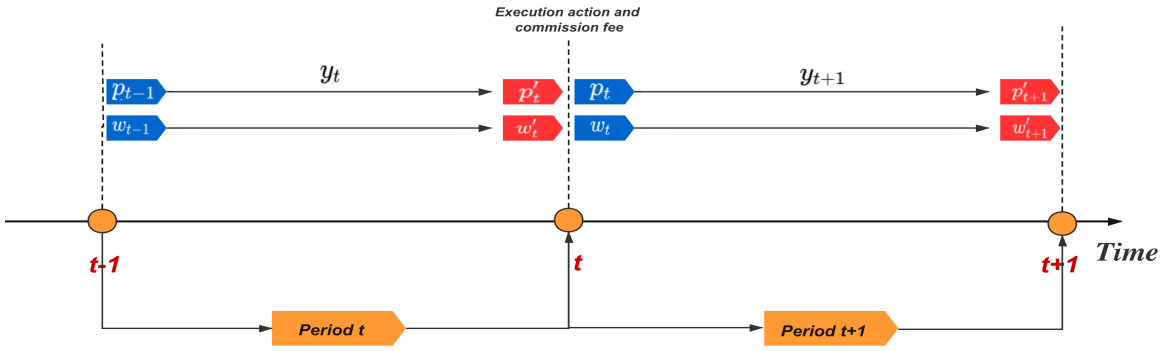


Figure 2.1: The structure of our trading system.

2.3.3 Assumption and restrictions

To simulate real-world market trades, we make several assumptions to formulate the problem. First of all, the actions are only executed at the end of the period. Second, we assume that the opening price is equal to the closing price of the previous day. After-sales market transactions are not allowed. Third, short selling is not allowed in our trading environment. Finally, we also assume the market is sufficiently liquid such that any transactions can be executed immediately with minimal market impact.

2.4 The Classical DDPG Algorithm

Portfolio management is a financial decision-making task, which aims at boosting the total profits or returns and lowering the risk via asset allocation. The asset allocation process can be constantly changed; therefore, we employ an off-policy agent using a DRL algorithm that maps the high dimensional state space to a high dimensional continuous action space. DDPG is an actor-critic based deep RL algorithm proposed in Lillicrap et al. (2015). It uses a neural network as a Q-function approximator and proposes a replay buffer to improve convergence to the optimal policy, because the proposed replay buffer resolves the problem that the learned action function is relatively unstable.

The classical DDPG algorithm has been developed by a Markov decision process, which consists of a state space $\mathcal{S}$, action space $\mathcal{A}$, an initial state distribution $p(s_0)$, transition dynamics $p(s_{t+1}|s_t, a_t)$, and reward function $r(s_t, a_t)$. The DDPG algorithm includes four

neural networks: the actor network, the critic network and their respective target networks. In the initial stage, we randomly initialize each network and reset the replay buffer, and the DDPG agent aims to learn from interaction with the environment. At the beginning of the training process, the current state, next state, action and the immediate reward from the environment are stored in a replay buffer, then DDPG assembles a mini-batch from the replay buffer and feeds it to both the actor, critic, and their target networks. Based on the sample mini-batched from replay buffer, the target actor network produces a target action according to the next state; the target Q-value is generated by the target critic network associated with the next states and target action. The target Q-values of the current actions and states are calculated from the immediate rewards and the discounted Q-values for the next states via the Bellman equation. The critic network is updated by minimizing the TD-error, calculated as the difference between target Q-value and actual Q-value; the actor network is trained by adopting the policy gradient for the critic network. Finally, the target network weights are updated using a soft updates strategy from actor and critic networks. A soft update strategy includes smoothly mixing the regular network weights with target network weights. The structure of classical DDPG is shown in Figure 2.2.

For the portfolio allocation problem, at each trading time t, we assume the DDPG agent only observes the market information of OHLCV data. With such an assumption, the observation s_t can be expressed as:

$$s_t = \Big[V_{1,t}, V_{2,t}, ..., V_{n,t}\Big], \tag{2.16}$$

where $V_{i,t}$ is defined by (2.2). We take the portfolio weight vector as an action, so action vector a_t is equal to weight vector w_{t-1}, where w_{t-1} denotes the portfolio weight vector at the beginning of period t. The DDPG agent aims to maximize the total profit, which is equivalent to maximizing the logarithmic return. Therefore, the reward function $r(s_t, a_t)$ taking into account the transaction cost is defined as

$$r(s_t, a_t) = \log \frac{p_t}{p_{t-1}} = \log((1 - C_t)y_t \cdot w_{t-1}). \tag{2.17}$$

Thus, we have the immediate reward at each timestep that avoids the sparsity of the reward problem. At each timestep t, the agent takes an action a_t based on the current observation s_t, and receives a reward $r(s_t, a_t)$. The total discounted future rewards until timestep T is

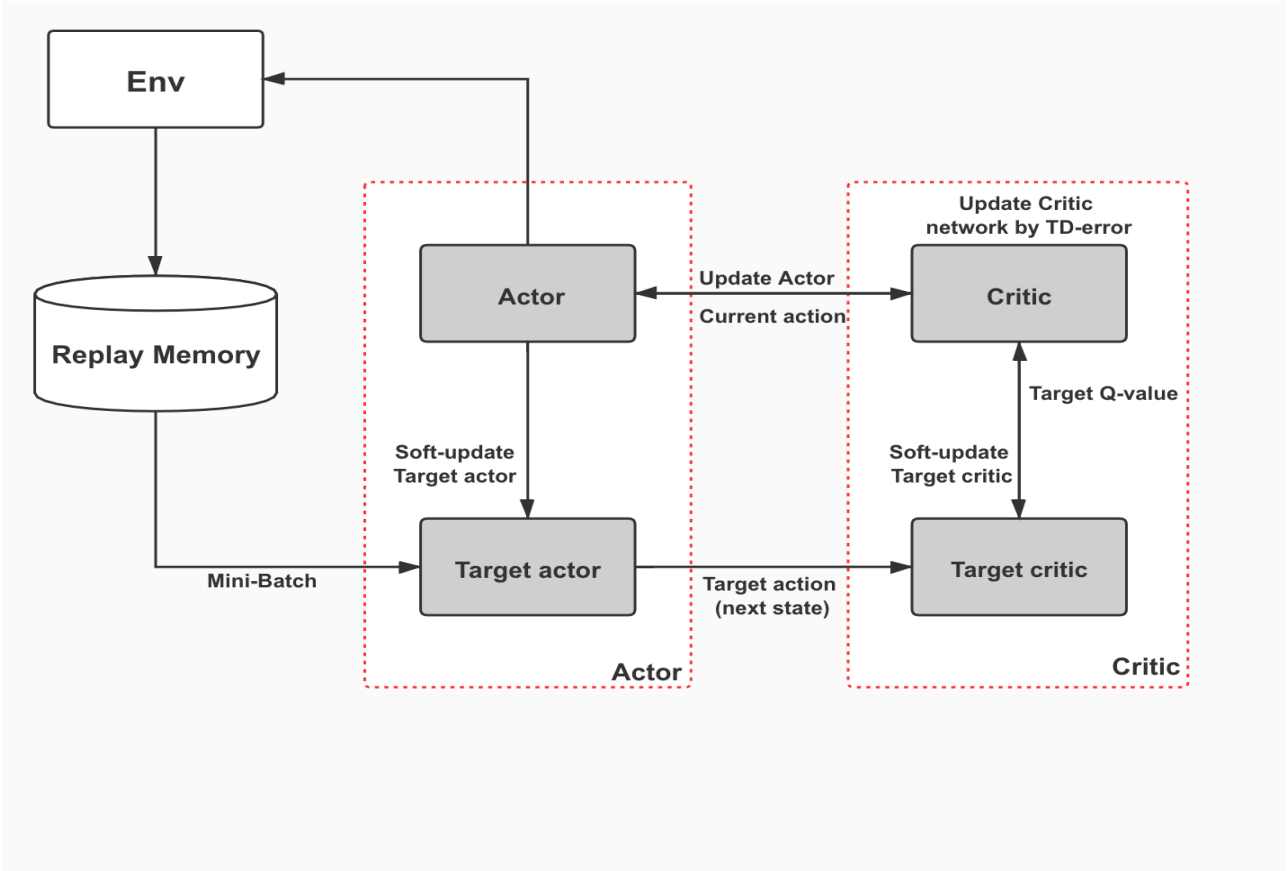


Figure 2.2: The structure of classical DDPG

given by

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i), \tag{2.18}$$

where the discount factor $\gamma \in [0, 1]$. The objective of reinforcement learning is to learn a policy by maximizing the expected discounted future rewards given the current state

$$J = \mathbb{E}\Big[R_t \Big| s_t\Big]. \tag{2.19}$$

By the Bellman equation, it allows us to compute the Q-value by recursion:

$$Q^{\mu}(s_t, a_t) = \mathbb{E}\Big[r(s_t, a_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1}))\Big]. \tag{2.20}$$

The parametrized actor function $\mu(s|\theta^{\mu})$ specifies the current policy by deterministically mapping states to a specific action $\mu : \mathcal{S} \rightarrow \mathcal{A}$. The critic network $Q(s, a)$ is updated by minimizing a squared TD-error below:

$$L = \frac{1}{N} \sum_{i} \Big[y_i - Q(s_i, a_i|\theta^{Q})\Big]^2, \tag{2.21}$$

where $y_i = r(s_i, a_i) + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ and N is the number of transitions mini-batched from replay buffer.

Note that y_i is calculated by a separate target network which is softly updated, $Q'(\cdot)$ and $\mu'(\cdot)$ represent the target critic and actor network with the parameter $\theta^{Q'}$ and $\theta^{\mu'}$, respectively. The actor is updated by the following gradient of J with respect to the parameter θ^{μ} based on the policy gradient theory from Silver et al. (2014)

$$\begin{aligned} \nabla_{\theta^{\mu}} J &= \mathbb{E}\Big[\nabla_{\theta^{\mu}} Q(s, a\Big|\theta^{Q})\Big|_{s=s_t, a=\mu(s_t|\theta^{Q})}\Big] \\ &= \mathbb{E}\Big[\nabla_a Q(s, a\Big|\theta^{Q})\Big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^{\mu}} \mu(s|\theta^{\mu})\Big|_{s=s_t}\Big]. \end{aligned} \tag{2.22}$$

2.5 Proposed approaches

2.5.1 The Distributional DDPG Model

Although most deep RL aims to optimize the decision-making rule in terms of the expected future discounted rewards, the agent sometimes for some specific purpose aims to seek a big win on rare occasions or avoid a rare likelihood of suffering a huge loss. To reduce

the effects of rare bad events, the distributional RL is proposed by Dearden, Friedman, and Russell (1998) and Engel, Mannor, and Meir (2005), which aims to adopt a Gaussian distribution to approximate the distribution of future returns and model the uncertainty under this approximate distribution. Distributional RL has the benefit of considering the risks that may exist when future returns are stochastic since the observable stat state can not capture the intrinsic randomness of the environment. In addition, if there is a high variance or heavy tail in return distribution, the strategy of maximizing average return may lead to over-estimation of the expected future reward. Motivated by Tang et al. (2020), Barth-Maron et al. (2018), and Bellemare, Dabney, and Munos (2017), we apply the distributional RL algorithm to the portfolio management problem, namely Distributional DDPG, which constructs a risk-sensitive policy to reduce the effects of disaster events or potential losses.

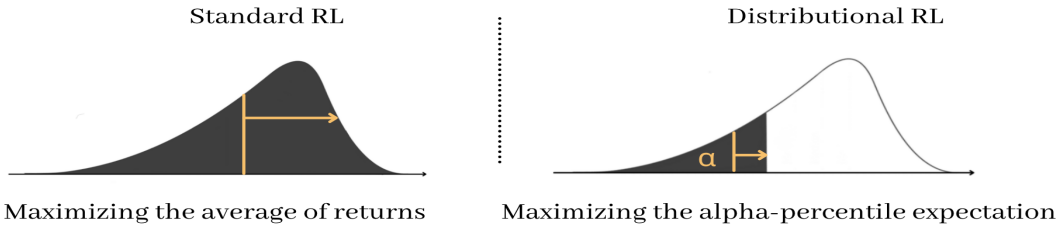


Figure 2.3: Distributional RL

The standard Markov decision process (MDP) consists of a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ present the continuous state and action space respectively, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function, γ denotes the discount factor, and $\mathcal{P}$ is the transition probability density of moving the current state into the next state. Suppose that R is a random variable for future return, $P(R|s, a)$ is the probability distribution of future returns, which is given the current state s and action a. The *α-percentile* expectation[3] that represents the expected return under the bottom *α-percentile* of the distribution over returns is employed as the criterion of distributional RL. The objective function can be formulated as:

$$J_\alpha = \mathbb{E}[R \big| R \leq \textit{percentile}(\alpha), s], \tag{2.23}$$

[3]We note that some literature call this CVaR, but we do not use it here to avoid confusion.

where $\alpha \in [0, 1]$ is a risk parameter. When $\alpha \to 0$, the strategy will concentrate on doing well in the worst case, while when $\alpha \to 1$, the strategy aims to perform well in the average performance. We combine the distributional RL with classical DDPG so that the critic learns to model the distribution over the expected total discounted reward. Our proposed method combines the distributional RL and DDPG, which enables critics to simulate the distribution of total discount rewards. As long as a good distribution can be learned by critics, then the actor network is updated by backpropagating the gradient back through the critic network.

Distributional DDPG includes an actor-critic network structure of the DDPG algorithm, and contains a distribution of future return $Z(s, a)$ that is a mapping from state-action pairs to distributions over returns. The distributional Bellman equation points out the distribution of $Z(s, a)$ is evaluated by three associated random variables: the reward r, the next state-action (s', a'), and its return $Z(s', a')$. The nature of distributional return $Z(s, a)$ is described by a following recursive equation:

$$Z(s, a) \stackrel{D}{=} r(s, a) + \gamma Z(s', a'), \tag{2.24}$$

where $U \stackrel{D}{=} V$ indicates that the random variable U has the same distribution pattern as V. In here, $Z(s, a)$ represents the inherent stochasticity of the interaction between the agent and the environment. Then, the transition operator P^{μ} is defined as:

$$P^{\mu} Z(s, a) \stackrel{D}{=} Z(s', a'), \quad s' \sim P(\cdot|s, a), a' \sim \mu(\cdot|s), \tag{2.25}$$

and the distributional Bellman operator $\mathcal{T}^{\mu}$ is given by

$$\mathcal{T}^{\mu} Z(s, a) \stackrel{D}{=} r(s, a) + \gamma P^{\mu} Z(s, a). \tag{2.26}$$

This implies that the two sides of a distributional equation relate to the distribution of two independent random variables, and it can be used to train the distributional reinforcement learning in many areas of research. Similar to the expected Bellman operator, the distributional Bellman operator can be proved to converge to the true return distribution. The convergence theory of the distributional Bellman operator has been proven by a contraction lemma, which needs to evaluate the distance between two return distributions (Rowland, Bellemare, Dabney, Munos, & Teh, 2018).

The Wasserstein metric is the main tool to measure the distance between cumulative distribution functions, proposed by Bickel and Freedman (1981). Different from the Kullback-Leibler (KL) divergence, the Wasserstein metric is a true probability that takes into account the probability of distances between various outcome events, which leads to the Wasserstein metric being well-suited for the field that exists an underlying similarity. For $p < \infty$, the p-th Wasserstein distance between two probability distributions F_U and F_V is defined as (Olkin & Pukelsheim, 1982):

$$W_p(U, V) = \Big(\int_0^1 \Big| F_U^{-1}(s) - F_V^{-1}(s) \Big|^p ds \Big)^{1/p}, \tag{2.27}$$

where F^{-1} is the inverse cumulative distribution function (CDF). Assuming that $U \sim \mathcal{N}(\mu_1, C_1)$ and $V \sim \mathcal{N}(\mu_2, C_2)$, the 2-Wasserstein distance simplifies to:

$$W_2(U, V) = |\mu_1 - \mu_2|^2 + C_1 + C_2 - 2(C_1 C_2)^{\frac{1}{2}}. \tag{2.28}$$

As in Tang et al. (2020), we model $Z(s, a)$ as a Gaussian distribution, which provides a closed-form of the α-*percentile* expectation[4]. The output of the critic network can be expressed as the estimated mean and variance of future returns $Z(s, a)$ with weights θ^Q:

$$f_{critic}(s, a, \alpha | \theta^Q) \rightarrow \{\hat{Q}(s, a, \alpha), \hat{\mathcal{V}}(s, a, \alpha)\}, \tag{2.29}$$

where $f_{critic}(s, a, \alpha|\theta^Q)$ denotes the critic network with input state s, action a, and risk parameter α. We adopt Convolutional Neural Network (CNN) for the critic network. Three hidden convolution layers with *Relu* activation function are added following the input layer. Then, we modify the output layer that predicts the estimated value of mean and variance of the future returns. The *Softplus* activation function is applied to predict the variance of the future returns, which keeps the variance always positive. The convergence proofs[5] for the critic network are given in Appendix A.1.

With the benefit of the critic's structure, the estimated $\hat{Q}(s, a, \alpha)$ and $\hat{\mathcal{V}}(s, a, \alpha)$ are applied to calculate the α-*percentile* expectation in closed-form. Let $\Gamma^\mu(s, a, \alpha)$ denotes the

[4]Without this assumption, it requires choosing another algorithm to approximate the sample Bellman updates and minimize the Wasserstein metric in each step, which is computationally too expensive.

[5]We note that the 2-Wasserstein distance W_2 cannot be directly used to bound the variance difference.

α-*percentile* expectation when the state s and executing action a, following policy μ hereafter, the closed-form of α-*percentile* expectation is formulated as:

$$\Gamma^{\mu}(s,a,\alpha) = \mathbb{E}[R|R \leq \textit{percentile}(\alpha), s, a] = \hat{Q}(s,a,\alpha) - \frac{\varphi(\alpha)}{\Phi(\alpha)}\sqrt{\hat{\mathcal{V}}(s,a,\alpha)}, \tag{2.30}$$

where $\varphi(\cdot) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is the standard normal p.d.f., and $\Phi(\cdot)$ is its CDF.
Therefore, the objective function (2.23) can be rewritten as:

$$J_{\alpha} = \mathbb{E}[R|R \leq \textit{percentile}(\alpha), s] = \int_S \rho^{\mu}(s) \int_A \mu_{\theta_a}(a|s,\alpha)\Gamma^{\mu}(s,a,\alpha)dads, \tag{2.31}$$

where ρ^{μ} denotes the stationary distribution over the state space given the policy μ.
Then, the actor is updated by the following deterministic gradient, which adopts the chain rule to the α-*percentile* expected return with respect to the actor parameters (Silver et al., 2014):

$$\nabla_{\theta_a} J_{\alpha} = \mathbb{E}\Big[\nabla_{\theta_a}\mu(a|s,\alpha)\hat{Q}(s,a,\alpha) - \frac{\varphi(\alpha)}{\Phi(\alpha)}\nabla_a\sqrt{\hat{\mathcal{V}}(s,a,\alpha)}\nabla_{\theta_a}\mu(a|s,\alpha)\Big]. \tag{2.32}$$

Note that the objective function J_{α} is dependent on the risk levels (αs). In the training process, we uniformly sample $\alpha \sim \textit{Uniform}(0,1)$ at the beginning of the episode and fix α for the whole episode. During the testing period, the policy μ can yield different actions in given the same state s, conditioned on the setting of α. Intuitively, a small value of α leads to conservative behavior while a larger value of α leads to more aggressive behavior (see Algorithm 1). The structure of Distributional DDPG is displayed in Figure 2.4.

2.5.2 The Hierarchical DDPG Model

In this subsection, we propose a novel algorithm, called Hierarchical DDPG, which adds the Hierarchical structure to the DDPG algorithm. Original Hierarchical RL refers to the concept of decomposing RL problem into sub-problems (sub-tasks). Solving each sub-task will be more vigorous and efficient than solving the whole problem. The investor's goal is to select the best portfolio with the highest total profit and lowest portfolio risk for his/her investment. However, when there is a chance of gains and losses, most investors would prefer to avoid losses. Our Hierarchical DDPG algorithm utilizes the structure of Hierarchical RL

Algorithm 1 Distributional DDPG

procedure TRAINING
Randomly initialize critic and actor network of agent with weights θ^Q and θ^μ.
Initialize target actor network and critic network with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$.
Initialize replay buffer $\mathcal{B}$
    for $episode = 1, M$ **do**
        Initialize an OU random process $\mathcal{N}$ for action exploration
        Receive initial observation state s_1
        Sample $\alpha \sim Uniform(0, 1)$
        for $t = 1, T$ **do**
            Sample action $a_t = \mu(s_t, \alpha|\theta^\mu) + \mathcal{N}$
            Observe reward r_t, next state s_{t+1} from environment
            Store transition $\{s_t, a_t, r_t, s_{t+1}, \alpha\}$ into $\mathcal{B}$
            Sample a random mini-batch of N transitions from $\mathcal{B}$
            Using target network to approximate $\mathcal{T}^\mu Z(s, a)$ distribution by calculating the mean and variance from the critic network.
            Update critic network θ^Q by minimizing Wasserstein distance in equation (2.28)
            Update actor network θ^μ by using sample deterministic policy in equation (2.32)
            Update the target network by soft-update
            $\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$
            $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$
        end for
    end for
end procedure

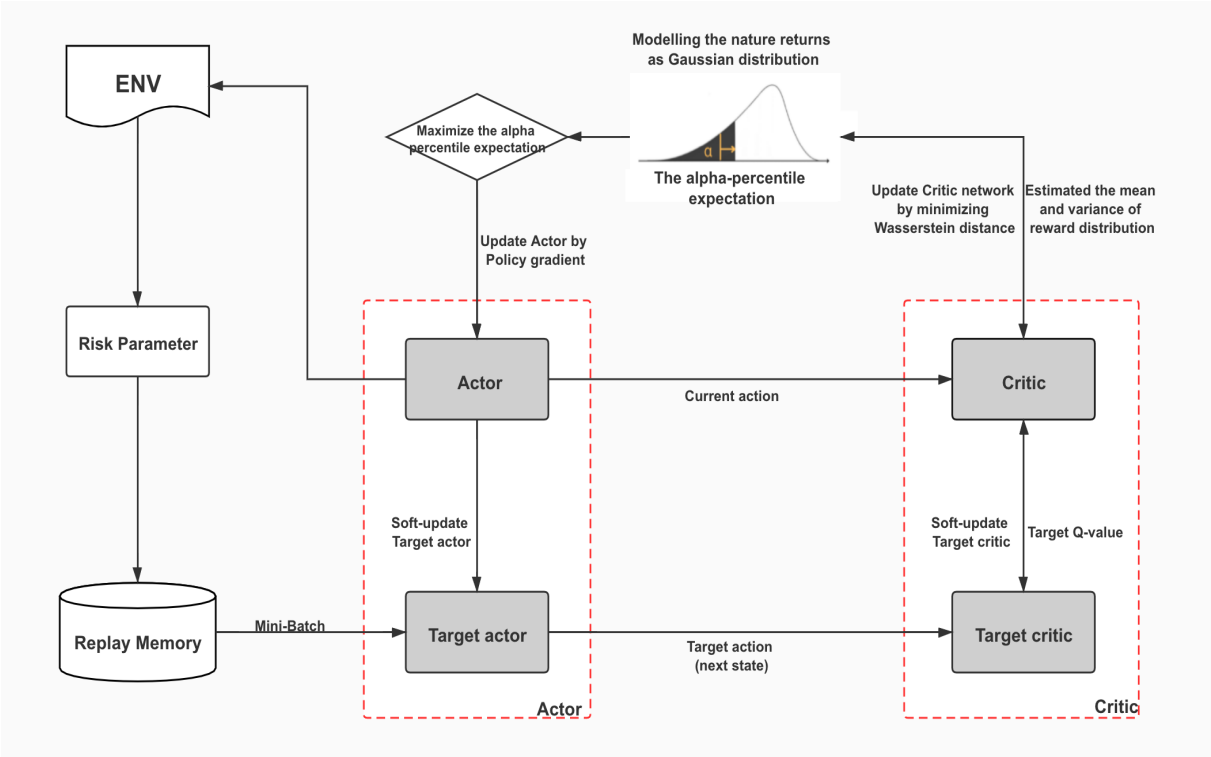


Figure 2.4: The structure of Distributional DDPG

in which the two-level mechanism allows the agent to avoid the potential loss, and balance the portfolio profit and risk for different scenarios.

The Hierarchical DDPG framework develops from classical DDPG and Hierarchical RL, which consists of lower-level and higher-level policy. The lower-level policy is an actor-critic based structure, and it is interpreted as a *worker* that selects primitive actions at every time step by maximizing the logarithmic rate of return when the portfolio risk is lower than a certain level. The higher-level policy is also an actor-critic based structure, and it is interpreted as a *manager* that selects the action according to the observation state and the action generated from the lower-level policy by minimizing the portfolio risk when the portfolio risk exceeds the tolerance of investors. In other words, the *manager* makes an adjustment to reduce the portfolio risk based on the *worker*'s action when the portfolio risk exceeds a certain level of risk. The critics of the *manager* and *worker* are used to evaluate their works. More specifically, by exploiting the market information from the environment, the *worker* observes the state from the environment, and produces the action g_t to maximize

the total profit. Then, we employ an indicator to check whether the portfolio risk exceeds the investor's tolerance. If the investor can afford the potential loss, then the action g_t will be executed and received the rewards from the environment. If not, the *manager* will adjust the *worker*'s trading strategy and yield the action a_t based on the observation state and the *worker*'s action g_t to reduce the portfolio risk. At the final stage, the *manager* will execute the action a_t in the environment. The main idea of Hierarchical DDPG is displayed in Figure 2.5.

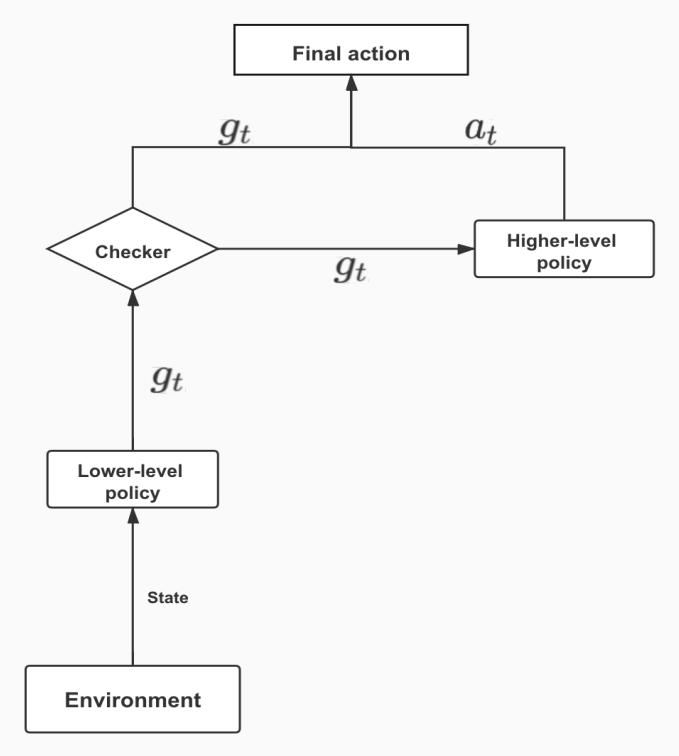


Figure 2.5: The main idea of Hierarchical DDPG

Now, we introduce the indicator to measure the portfolio risk. Conditional Value-at-Risk (CVaR) is often used as a measure of risk and is also referred to as expected excess loss or expected shortfall. CVaR is a coherent risk measure and more attractive compared to Value-at-Risk (VaR) because it takes into account the contribution from the very rare but very large losses. Rockafellar, Uryasev, et al. (2000) employ CVaR as the risk measure and minimize CVaR to compute an optimal investment portfolio. Krokhmal, Palmquist, and Uryasev (2002) propose a new approach for optimizing CVaR in portfolio optimization problems.

They extend the Rockafellar et al. (2000) approach by maximizing expected returns under CVaR constraints. CVaR constraints are used to limit the percentiles of the loss distribution and sculpt the loss distribution according to the decision makers' preferences. Linear Programming (LP) approach is one of the standard approaches for solving CVaR optimization problems. A piecewise linear function can approximate the typical continuously differentiable CVaR function by adopting the Monte Carlo simulation. In this chapter, we apply the parametric CVaR approach under the Gaussian distribution to measure the portfolio risk. Parametric CVaR_α can be formulated as a closed-form:

$$\text{CVaR}_\alpha(a_t) = \mathscr{V}_t \frac{\varphi(\Phi^{-1}(\alpha))}{\alpha} - \mu_t, \tag{2.33}$$

where μ_t and $\mathscr{V}_t$ are the mean and variance of the portfolio return defined in (2.8), and $\varphi(\cdot)$ is the standard normal p.d.f., $\Phi(\cdot)$ is the standard normal CDF, so $\Phi^{-1}(\alpha)$ is the standard normal quantile. When the portfolio risk is below the CVaR risk constraints ($\text{CVaR}_\alpha \leq C$), the lower-level policy aims to seek an aggressive trading strategy by maximizing the total profit (logarithm rate of return). On the other hand, when it exceeds the CVaR constraints ($\text{CVaR}_\alpha > C$), the main goal is to reduce the portfolio risk instead of maximizing the total profit. The higher-level policy makes an adjustment of trading strategy and aims to seek a conservative trading strategy to reduce the risk by maximizing the expected future discount reward of the higher-level policy. The structure of Hierarchical DDPG is shown in Figure 2.6.

Next we introduce the reward of the lower-level policy and higher-lever policy. The reward of the lower-level policy is the same as in the classical DDPG algorithm (see (2.17)), and the reward function for higher-level policy is defined as

$$r'(s_t, g_t, a_t) = \text{CVaR}_\alpha(g_t) - \text{CVaR}_\alpha(a_t). \tag{2.34}$$

This implies the main goal of higher-level policy is to reduce the portfolio risk immediately compared with the lower-level's action. Then, the objective function for higher-level policy is given by

$$J^{(H)} = \mathbb{E}[R'_t | s_t], \tag{2.35}$$

where $R'_t = \sum_{i=t}^{T} \gamma^{i-t} r'(s_i, g_i, a_i)$.

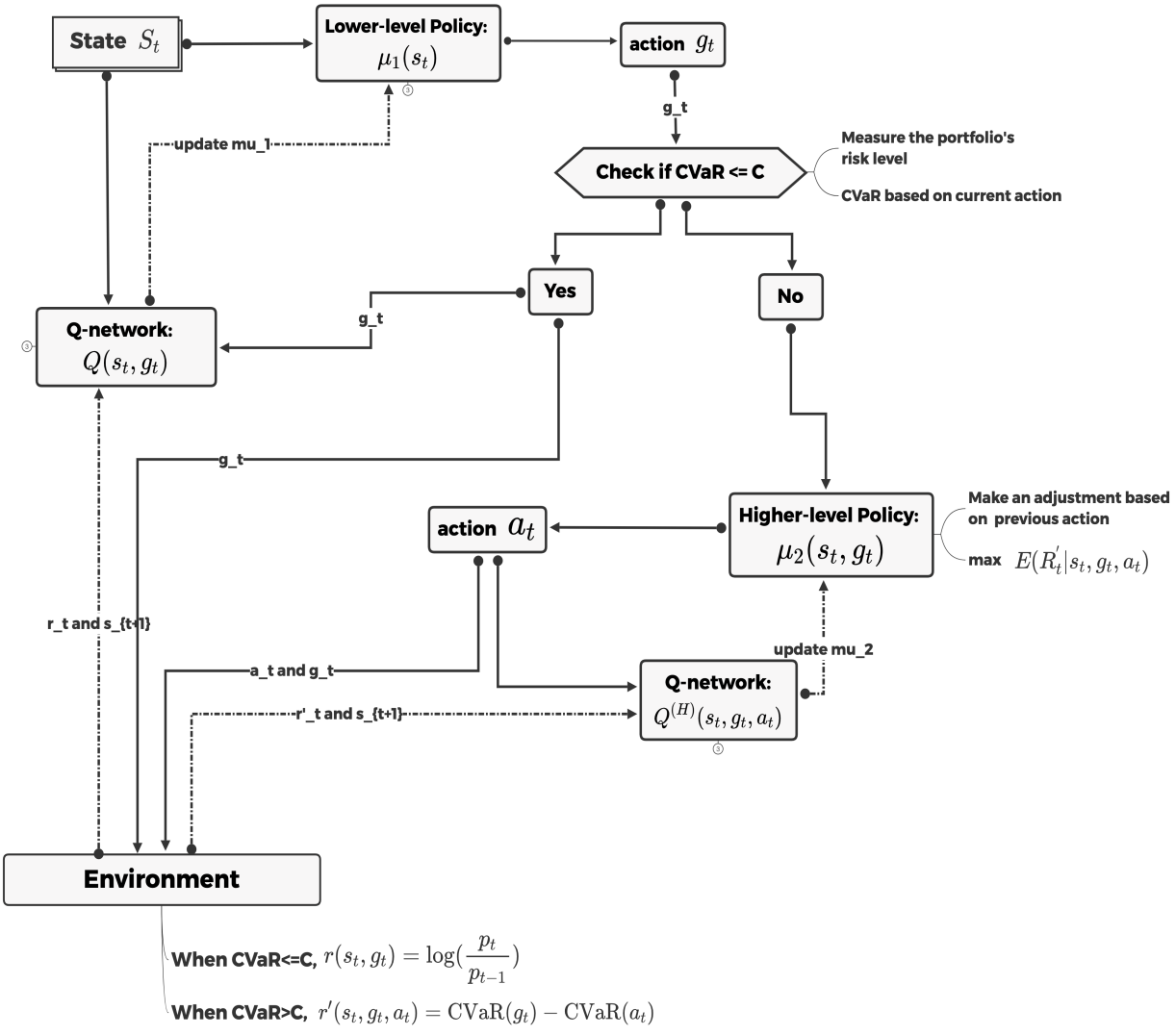


Figure 2.6: The structure of Hierarchical DDPG

Define the $Q^{(H)}$ function for higher-level network as:

$$Q^{(H)}(s_t, g_t, a_t) = \mathbb{E}[R'_t | s_t, g_t, a_t], \tag{2.36}$$

where g_t is the action from the lower-level policy, and a_t is the action from the higher-level policy.

Applying the Bellman equation to (2.36), we have

$$Q^{(H)}(s_t, g_t, a_t) = r(s_t, g_t, a_t) + \gamma Q^{(H)}(s_{t+1}, \mu_1(s_{t+1}), \mu_2(s_{t+1}, \mu_1(s_{t+1}))), \tag{2.37}$$

where $\mu_1(\cdot)$ is the policy from the lower-level, and $\mu_2(\cdot, \cdot)$ is the policy from the higher-level. To update the higher-level policy network, the policy gradient with respect to the parameter θ^{μ_2} is given by

$$\begin{aligned}\nabla_{\theta^{\mu_2}} J^{(H)} &= \mathbb{E}_{s_t}\left[\nabla_{\theta^{\mu_2}} Q^{(H)}(s, g, a|\theta^H)\Big|_{s=s_t, g=\mu_1(s_t), a=\mu_2(s_t, \mu_1(s_t)|\theta^{\mu_2})}\right] \\ &= \mathbb{E}_{s_t}\left[\nabla_a Q^{(H)}(s, g, a|\theta^H)\Big|_{s=s_t, g=\mu_1(s_t), a=\mu_2(s_t, \mu_1(s_t)|\theta^{\mu_2})} \nabla_{\theta^{\mu_2}} \mu_2(s, g|\theta^{\mu_2})\Big|_{s=s_t, g=\mu_1(s_t)}\right].\end{aligned} \tag{2.38}$$

This is derived in the same way as (2.22). The Hierarchical DDPG algorithm is given below (see Algorithm 2).

2.6 Experiment Results

2.6.1 Data

We conduct various experiments to verify our proposed approaches by using four different index ETFs: "SPY", "VGK", "GXC", and "EWG". SPY is the S&P 500 index ETF, which measures the stock performance of 500 large companies in the U.S. Market. VGK is the index ETF for the European All Cap developed by FTSE, which tracks the performance of major markets in Europe. GXC is one of the most comprehensive China equity funds available to U.S. investors, which is dominated by holding large-cap stocks and delivers greater diversification from a security perspective. Lastly, EWG aims to provide concentrated exposure to large and midcap segments of the German equity market, meaning it covers the top 85% of the German companies by market cap. It primarily consists of stocks traded on

Algorithm 2 Hierarchical DDPG

procedure TRAINING
Randomly initialize the critic and the actor networks of agent with weights θ^Q and θ^{μ_1}.
Randomly initialize the higher-level actor and critic networks with weights θ^H and θ^{μ_2}.
Initialize target networks and critic networks with weights $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{\mu_1'} \leftarrow \theta^{\mu_1}$.
Initialize replay buffer $\mathcal{B}_1$ and $\mathcal{B}_2$.
    for *episode* $= 1, M$ **do**
        Initialize an OU random process $\mathcal{N}$ for action exploration
        Receive initial observation state s_1
        for $t = 1, T$ **do**
            Sample action $g_t = \mu_1(s_t|\theta^\mu) + \mathcal{N}$
            Check the portfolio's risk level
            if CVaR $\leq$ C **then**
                Observe reward r_t, next state s_{t+1} from environment
                Store transition $\{s_t, g_t, r_t, s_{t+1}\}$ into $\mathcal{B}_1$
                Sample a random mini-batch of N_1 transitions from $\mathcal{B}_1$
                Set $y_i = r(s_i, g_i) + \gamma Q'(s_{i+1}, \mu_1'(s_{i+1}|\theta^{\mu_1'}))$ for all $i \in N_1$
                Update θ^Q by minimizing loss $L(\theta^Q) = \frac{1}{N}\sum_i (y_i - Q(s_i, g_i|\theta^Q))^2$
                Update the actor policy θ^{μ_1} using the sampled policy gradient:

$$\frac{1}{N_1}\sum_i \nabla_g Q(s, g|\theta^Q)|_{s=s_i, g=\mu_1(s_i)} \nabla_{\theta^{\mu_1}} \mu_1(s|\theta^{\mu_1})|_{s=s_i}. \tag{2.39}$$

                Update the target network by soft-update
                $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$.
                $\theta^{\mu_1'} \leftarrow \tau\theta^{\mu_1} + (1-\tau)\theta^{\mu_1'}$.
            end if
            if CVaR $>$ C **then**
                $a_t = \mu_2(s_t, g_t, a_t)$
                Store transition $\{s_t, g_t, a_t, r_t, s_{t+1}\}$ into $\mathcal{B}_2$
                Sample a random mini-batch of N_2 transitions from $\mathcal{B}_2$
                Set $y_j^{(H)} = r(s_j, g_j, a_j) + \gamma Q^{(H)}(s_{j+1}, \mu_1(s_{j+1}), \mu_2(s_{j+1}, \mu_1(s_{j+1})))$ for all $j \in N_2$
                Update θ^H by minimizing loss $L(\theta^H) = \frac{1}{N_2}\sum_j (y_j^{(H)} - Q^{(H)}(s_j, g_j, a_j|\theta^H))^2$
                Update the higher-level policy θ^{μ_2} using the sampled policy gradient:

$$\frac{1}{N_2}\sum_j \nabla_a Q^{(H)}(s, g, a|\theta^H)\Big|_{s=s_j, g=\mu_1(s_j), a=\mu_2(s_j, \mu_1(s_j))} \nabla_{\theta^{\mu_2}} \mu_2(s, g|\theta^{\mu_2})\Big|_{s=s_j, g=\mu_1(s_j)}. \tag{2.40}$$

            end if
        end for
    end for
end procedure

the Frankfurt Stock Exchange. The data set, obtained from Yahoo Finance, consists of daily prices and volume data over a 10-year period from 2010-01-01 to 2020-07-30. The training set and testing set are distributed according to the ratio of 8: 2.For the purpose of training and testing, two independent trading environments are designed. In addition, short selling is not allowed and a commission rate of 0.25% will be deducted for each transaction for all experiments.

Data Preprocessing

The absolute prices and volumes of the assets, i.e., opening, highest, lowest, closing prices, and volume in the problem are not sensitive to the agent for making any trading decisions, but the changes in prices and volumes are important to the agent. Therefore, the input prices and volumes to the network need to be normalized. To be specific, we divide the opening, closing, highest, lowest prices by the closing price on the last day of the period, and divide the volumes by the volume on the last day of the period. For example, the input state with window size m and number of assets n is given by

$$s_t = \left[V'_{1,t}, V'_{2,t}, ..., V'_{n,t}\right], \tag{2.41}$$

where $V'_{i,t}$ is the information of the i-th stock at time t after normalization, given by

$$V'_{i,t} = \begin{bmatrix} \frac{v^o_{i,t}}{v^c_{i,t}}, & \frac{v^h_{i,t}}{v^c_{i,t}}, & \frac{v^l_{i,t}}{v^c_{i,t}}, & 1, & 1 \\ \frac{v^o_{i,t-1}}{v^c_{i,t}}, & \frac{v^h_{i,t-1}}{v^c_{i,t}}, & \frac{v^l_{i,t-1}}{v^c_{i,t}}, & \frac{v^c_{i,t-1}}{v^c_{i,t}}, & \frac{v^v_{i,t-1}}{v^v_{i,t}} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \frac{v^o_{i,t-m+1}}{v^c_{i,t}}, & \frac{v^h_{i,t-m+1}}{v^c_{i,t}}, & \frac{v^l_{i,t-m+1}}{v^c_{i,t}}, & \frac{v^c_{m,t-m+1}}{v^c_{i,t}}, & \frac{v^v_{i,t-m+1}}{v^v_{i,t}} \end{bmatrix}. \tag{2.42}$$

2.6.2 Evaluation metrics

Portfolio optimization problems involve determining the best asset allocation for an investment fund in accordance with specific objectives, such as maximizing portfolio return and minimizing portfolio risk. Assessing the performance of a trading strategy objectively and rationally can be a challenging and difficult task. An outperforming trading strategy is not only expected to generate a higher profit but also alleviate the portfolio risk associated with

Table 2.2: Data Description

ETFs	Description
SPY	SPY is one of the most popular and oldest ETFs designed to track the Standard & Poor's 500 index. It holds a portfolio of 500 securities, which are selected by the S&P Committee to represent large-cap companies in the United States. At present, the top 3 sectors in which SPY holds shares are technology, finance and health care, and the top 5 stocks in the portfolio include Microsoft, Apple, Amazon, Facebook and Berkshire Hathaway.
VGK	VGK tracks all capitalization and market capitalization weighted indices of developed European securities. It is a subset of the FTSE global stock index series, covering about 98% of the global market, and diversified enough to invest across various industries.
GXC	GXC tracks the S&P China BMI, which is a rules-based index that measures the performance of global equity markets. It includes major share classes like A, B, H, red chips, P chips, and foreign listings. The fund typically invests almost all (but at least 80%) of its total assets in the securities comprising the index.
EWG	EWG tracks a market-cap-weighted index of large and midcap German companies. It aims to provide concentrated exposure to large- and midcap segments of the German equity market, meaning it covers the top 85% of the German companies by market cap. It primarily consists of stocks traded on the Frankfurt Stock Exchange.

the trading activity. In this subsection, three evaluation metrics are introduced to assess the portfolio performance, that is, *Accumulated return*, *Sharpe ratio*, and *Maximum drawdown*.

Accumulated return

The Accumulated return is one of the popular evaluation metrics used to assess the portfolio profit. The higher Accumulated return implies the portfolio yields a higher profit. Consider a portfolio having the arithmetic return R_t at time t, the accumulated return can be calculated as

$$\textit{Accumulated return} = \prod_{i=1}^{t}(1 + R_i). \tag{2.43}$$

This is the standard metric used to compare performance and relates the wealth at time t, W_t, with the initial wealth, W_0, as $W_t = W_0 \times \textit{Accumulated Return}$. In this chapter, all trading experiments adopt initial wealth $W_0 = 1$.

Sharpe ratio

The Sharp ratio is a performance metric that is widely and frequently used in the fields of finance and portfolio management because it takes into account both the profit and risk of the portfolio. This indicator is developed by Nobel laureate William F. Sharpe, and expresses as the excess return per unit of risk that is evaluated as the standard deviation of return. The Sharp ratio can be written as follows:

$$\textit{Sharpe ratio} = \frac{E(R_t) - R_f}{\sigma(R_t)}, \tag{2.44}$$

where R_f is a risk-free return, $E(R_t)$ and $\sigma(R_t)$ represent the expectation and standard deviation of returns, respectively.

Maximum drawdown

Maximum drawdown (MDD) is another metric to assess the potential loss that seeks the maximum change from the highest to the lowest. It is dedicated to capital preservation that is the main anxiety for most rational investors. For example, when MDD is quite small, it implies a minor loss from investment, and when an investment has never been lost, the MDD would be zero. On the other hand, the worst possible maximum drawdown would be -100%, meaning the investment is completely worthless. The maximum drawdown can be calculated as follows:

$$\textit{Max drawdown} = \max \frac{R_t - R_{t+1}}{R_t}, \tag{2.45}$$

where R_t and R_{t+1} represent the rate of return in period t and $t+1$, respectively.

2.6.3 The Results

This subsection presents the experimental results of our proposed methods and evaluates the effectiveness of our approaches. We obtain the opening, highest, lowest, closing prices, and volume of four ETFs. These four ETFs have different patterns, the movement of the closing prices and their details are described in Figure 2.7. These prices are expressed in the U.S. dollar. As shown in Figure 2.7, GXC shows the most gradual increase in these ETFs; SPY shows an upward trend but has been more volatile; EWG and VGK do not show a particular movement.

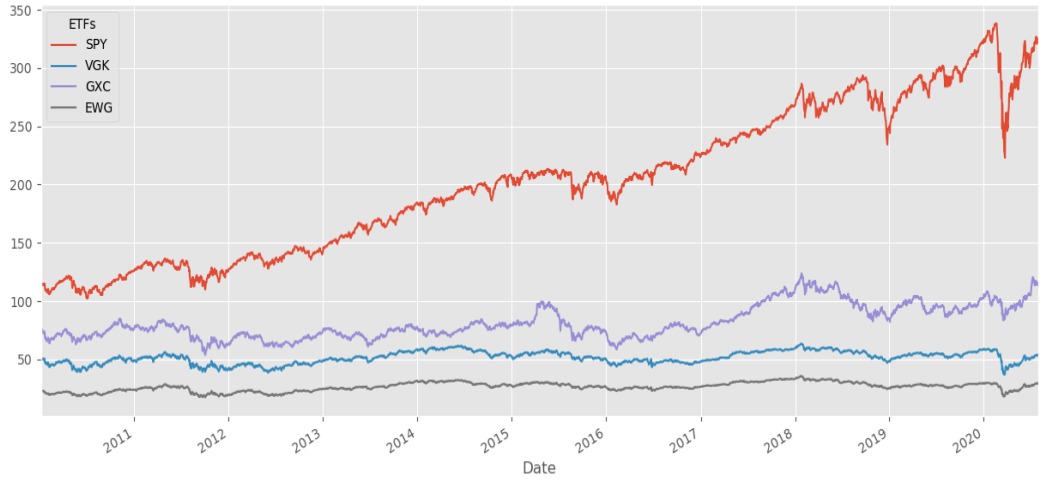


Figure 2.7: The closing prices of each ETF.

2.6.3.1 The experimental results for DDPG

The DDPG agent is trained on the training environment and tested on the testing environment separately. The window size of the trading system is ten trading days, which indicates that the DDPG agent can observe the prices and trading volumes in the past ten days. The training process is carried for 500 iterations, and each iteration consists of 128 steps until

the actor and critic networks get convergence to the optimal. To avoid convergence to local optimum policies, the weights of actor and critic are saved for the best performance. The actor and critic network adopts CNN with three hidden layers, and each convolution layer is a fully-connected layer with activation function of *ReLu*. The weights of actor and critic networks are randomly initialized at the beginning of each episode. The *Softmax* outputs of the actor network generate the actual corresponding portfolio weights.

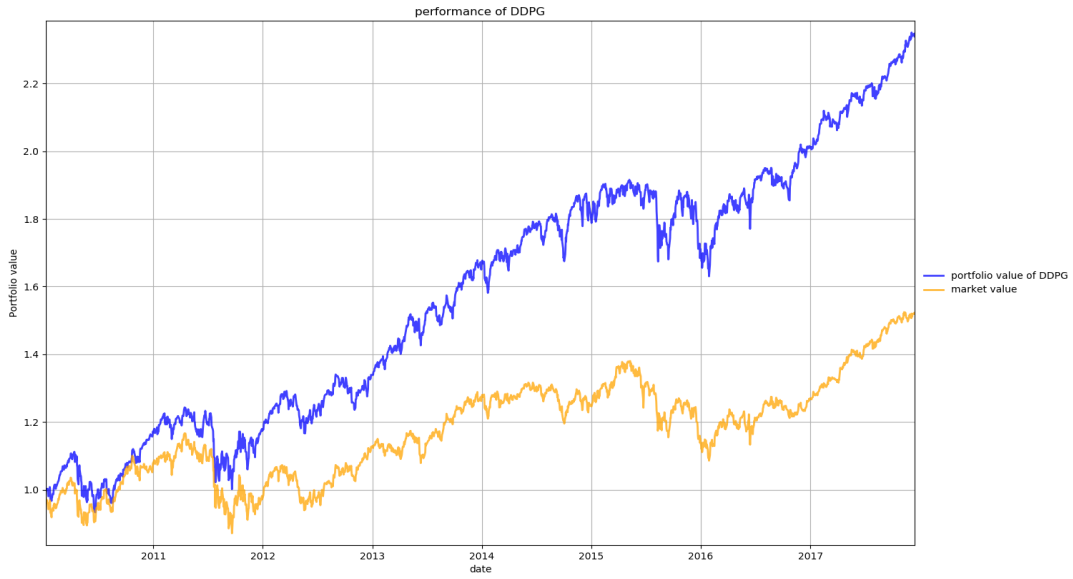


Figure 2.8: The portfolio value of DDPG under window size of ten-day during the training period.

Figure 2.8 and Figure 2.9 show the performance of DDPG with the ten-day window size for the training and testing period, respectively. As shown in Figure 2.8, the portfolio value has surprisingly increased by 138.84% during the training period; it achieves outstanding performance compared to the market value. Here, the market value presents a portfolio that consists of equally-weighted investment assets. The maximum drawdown and Sharpe ratio of the DDPG portfolio are 8.66% and 80.48%, respectively. In addition, Figure 2.9 shows that the portfolio of DDPG has given a 10.05% accumulated rate of return on investment at the end of the testing period, and their maximum drawdown and Sharpe ratio indices are 13.58% and 36.47%, respectively. These evidences indicate that the trading strategy of the

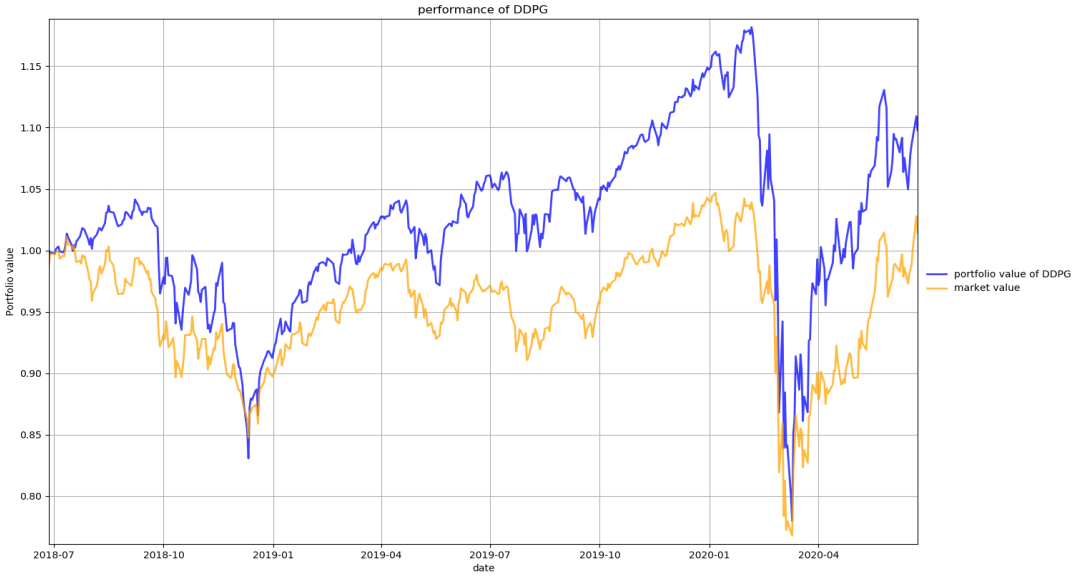


Figure 2.9: The portfolio value of DDPG under window size of ten-day during the testing period.

DDPG agent has a higher potential risk and suffers a massive loss when the financial market crashes.

In addition, we test the effect of window size on the portfolio performance. The different window sizes $(5, 10, 20, 25)$ are applied to our experiment, the experimental results during the testing period are displayed in Table 2.3. While using window sizes of 20 and 25, the accumulated return increases to 12.93% and 14.74%, improved by 2.43% and 4.24% compared to the case of ten-day window size, respectively. Furthermore, the Sharpe ratio rises to 38.44% and 39.94% at the end of the testing period compared to the case of ten-day window size. These imply that the DDPG agent can construct a better portfolio when she observes more trading prices and volumes. One possible explanation is that the DDPG agent can predicate more accurate trends or movements based on more information she observed. The portfolio performances with different window sizes for the training and testing period are presented in Appendix A.2.

2.6.3.2 The experimental results for Distributional DDPG

Figure 2.10 shows the price movements of the portfolio with different risk parameters α during the testing period with the ten-day window size. Although the accumulated portfolio value has not increased much, the maximum drawdown has significantly decreased. When $\alpha = 5\%$, the agent only takes into consideration the worst-case, the agent is willing to choose cash instead of investing funds in other risky assets. As α increases, the agent is not willing to consider the extreme cases, and aims to allocate more investment funds into these risky assets. Therefore, we can see from Figure 2.10 that the accumulated return increases to 4.22% and 10.32% when risk parameter α are 15% and 30%, respectively. Also, it reveals that the maximum drawdown of the distributional DDPG portfolio decreases as the risk parameter α decreases. This illustrates that the investor may suffer a larger potential loss during the financial crisis when the investor is willing to tolerate more risk. Furthermore, the Sharpe ratio increases as the α increases, which indicates that the earning per unit risk of this portfolio increases when the agent is willing to take more risk. These points demonstrate that Distributional DDPG can construct a more robust trading policy according to the investor's risk preference.

In addition, we test the effect of window size on the portfolio performance, and the experimental results of the portfolio performance with different window sizes and αs are shown in Table 2.3. No matter what the value of windows size is, the optimal trading strategies are very conservative when the agent only considers the worst-case scenarios. When the window size is large, the agent observes more information for decision-making. Thus, she can learn more accurate trends or price movements because the noise and uncertainty of the market can be significantly reduced. For instance, when the window size is 25, the agent is willing to allocate more investment funds into risky assets instead of only holding cash even for the extreme case of risk parameter $\alpha = 5\%$. Therefore, the accumulated return under window size of twenty-five-day is higher than in other window sizes when the risk parameter α is 5%.

Overall, these experimental results provide strong evidence to demonstrate that the distributional DDPG method is an effective and efficient way to avoid a huge potential loss.

However, we find it hard to balance the total profit and portfolio risk. If the agent only considers the worst-case scenario, the portfolio will lose the potential gains; on the other hand, if the agent is willing to take more risk, the agent has to face large possible losses.

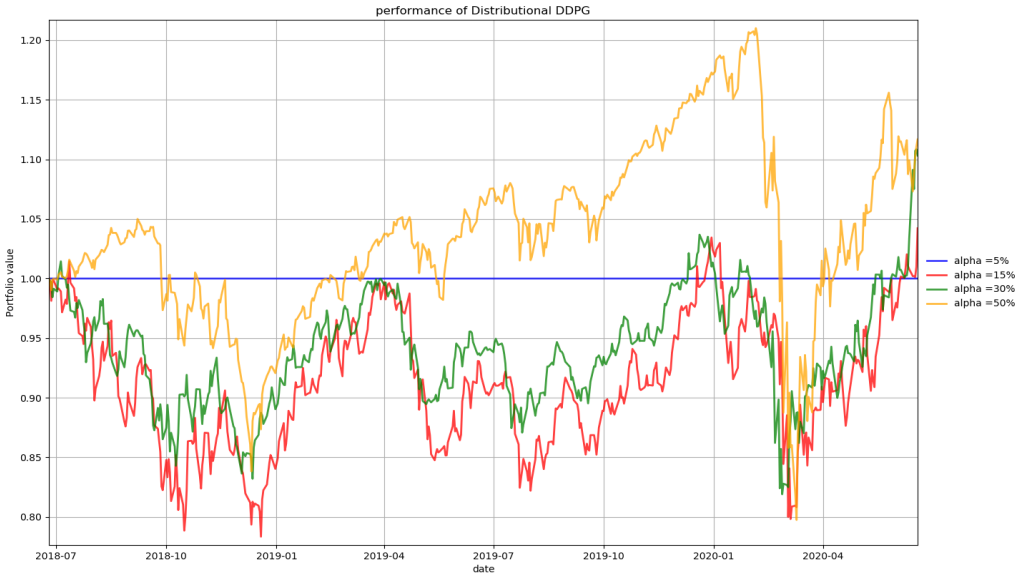


Figure 2.10: The portfolio value of Distributional DDPG with different risk parameters α under window size of ten-day.

2.6.3.3 The experimental results for Hierarchical DDPG

Figure 2.11 shows the Hierarchical DDPG portfolio with different CVaR constraints under the window size of ten-day. We obtain that the maximum drawdown has decreased in all cases, e.g., when the CVaR constraint is 5%, the maximum drawdown is 10.05%, reduced by 3.54% compared to classical DDPG. We observe that the accumulated rate of return and Sharpe ratio of Hierarchical DDPG have improved by 4.57% and 32.91%, respectively. Also, the case of the CVaR constraint $C = 13\%$ provides the most significant maximum drawdown of 12.95%, and the case of $C = 8\%$ provides the highest Sharpe ratio of 76.83%. These shreds of evidence illustrate that the Hierarchical DDPG model is superior to DDPG to avoid losses in the recession market. Table 2.3 shows the experimental results of Hierarchical DDPG

Figure 2.11: The portfolio value of Hierarchical DDPG with different constraints C under window size of ten-day.

with different CVaR constraints and window sizes. Compared to classical DDPG, it reveals that the Hierarchical DDPG algorithm performs better than the classical DDPG method in perspective of the maximum drawdown in most cases except the cases that window size is 20 or 25, and constraint C is 8%. The accumulated return and Sharpe ratio of Hierarchical DDPG are higher than those of classical DDPG in many cases. For example, when the window size is 5, the accumulated rates of return with different constraints $(5\%, 8\%, 13\%)$ have improved by $1.77\%, 6.2\%$, and 1.77%, respectively. These experimental results demonstrate that our proposed Hierarchical DDPG provide stable results with different window sizes and constraints. Overall, the Hierarchical DDPG agent can achieve higher rate of return and higher Sharpe ratio compared to classical DDPG, and moreover, control the short-term risk within a reasonable range.

In Figure 2.12, we compare the performance of classical DDPG and the proposed approaches during the testing period. For the window size of ten-day, we display the case of CVaR constraint $C = 5\%$ as an example to present the performance of Hierarchical DDPG, and the case of risk parameter $\alpha = 30\%$ as an example to present the performance

of Distributional DDPG. As shown in Figure 2.12, we can see that these three approaches outperform the market value. Hierarchical DDPG provides the highest accumulated return and the lowest maximum drawdown compared to classical DDPG and Distributional DDPG. Specifically, the maximum drawdown drops from 13.59% to 10.05%, and the accumulated return rises from 10.5% to 15.07%. It illustrates that Hierarchical DDPG is an effective approach to avoid a huge loss caused by the financial crisis. Obviously, Distributional DDPG has a lower maximum drawdown and a higher accumulated return compared to the classical DDPG method. On the other hand, Figure 2.13 shows the portfolio risk of our approaches. In this chapter, we apply parametric CVaR as a risk measure for evaluating portfolio risk. It shows that the portfolio risk of Hierarchical DDPG keeps lower than the CVaR constraint, and Distributional DDPG has a lower portfolio risk than classical DDPG.

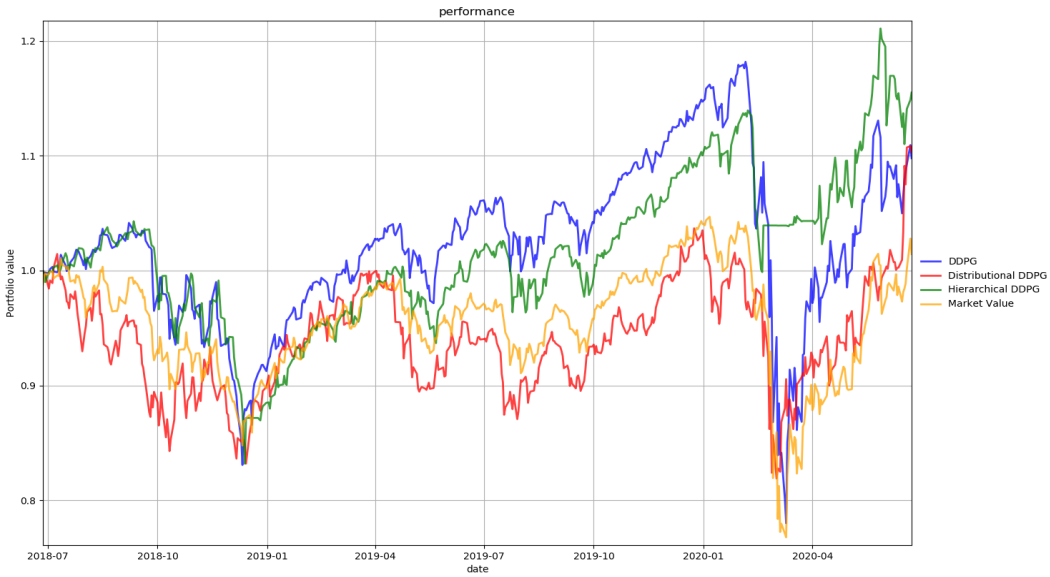


Figure 2.12: The performance comparison of Hierarchical DDPG, Distributional DDPG, and classical DDPG under window size of ten-day. Among them, Hierarchical DDPG represents the case of CVaR constraint $C = 5\%$, Distributional DDPG represents the case of risk parameter $\alpha = 30\%$.

In summary, the results demonstrate that Hierarchical DDPG and Distributional DDPG perform better than the classical DDPG algorithm for the rare occurrences of catastrophic

events, and they provide the capability to protect the investor who may suffer a massive loss in the recession market. Furthermore, Hierarchical DDPG appears to be a better approach to balance the portfolio risk and portfolio profit compared to Distributional DDPG, which provides a higher return and a lower portfolio risk or maximum drawdown.

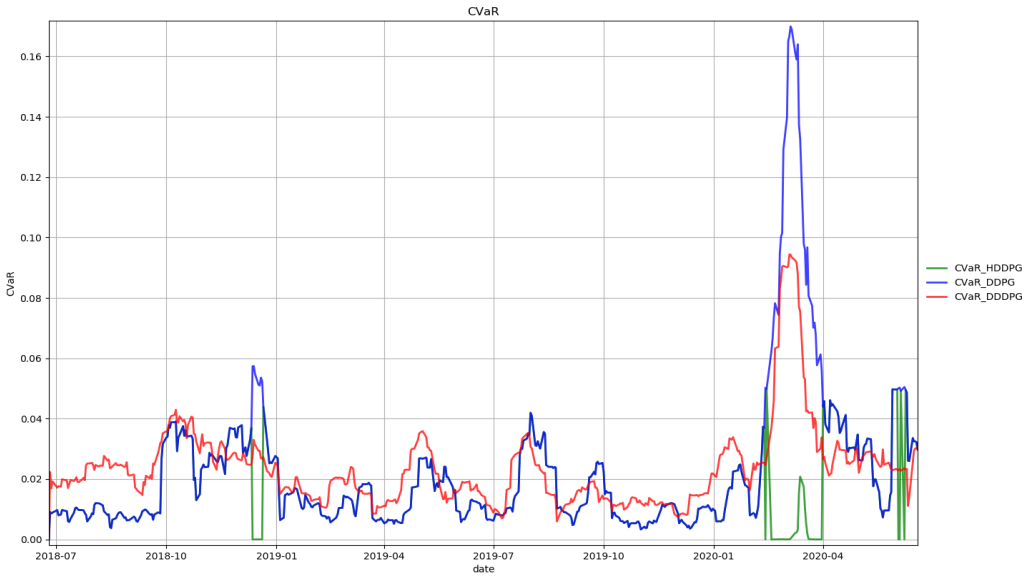


Figure 2.13: The portfolio risk comparison of Hierarchical DDPG, Distributional DDPG, and classical DDPG under window size of ten-day. Among them, Hierarchical DDPG represents the case of CVaR constraint $C = 5\%$, Distributional DDPG represents the case of risk parameter $\alpha = 30\%$.

2.6.3.4 Model validation with additional dataset

In this subsection, we validate our approaches by applying four different stocks from the U.S. stock market, that is, "AMZN", "CCL", "CVX", and "LUV". AMZN represents Amazon.com Inc, one of the world's largest e-commerce companies headquartered in Seattle, which focuses on cloud computing, digital streaming, and artificial intelligence. CCL stands for Carnival Corporation, the world's leading leisure travel company that offers extraordinary vacations to travellers around the world. CVX stands for Chevron Corporation, one of

Table 2.3: Experimental Results

	DDPG			Distributional DDPG				Hierarchical DDPG			
Window size	AR	MDD	SR	α	AR	MDD	SR	C	AR	MDD	SR
5	11.10%	13.59%	38.45%	5%	0.00%	0.00%	0.00%	5%	12.88%	10.30%	35.68%
				15%	2.41%	9.88%	17.71%	8%	17.31%	11.70%	51.24%
				30%	11.18%	13.59%	33.76%	13%	12.88%	11.70%	38.98%
				50%	12.70%	13.59%	36.52%				
10	10.50%	13.58%	36.47%	5%	0.00%	0.00%	0.00%	5%	15.07%	10.05%	69.38%
				15%	4.22%	6.88%	21.10%	8%	17.30%	11.68%	76.83%
				30%	10.32%	9.68%	34.05%	13%	16.00%	12.95%	58.14%
				50%	11.67%	13.59%	34.61%				
20	12.93%	13.59%	38.44%	5%	0.00%	0.00%	0.00%	5%	21.18%	10.30%	66.14%
				15%	8.37%	9.95%	28.58%	8%	2.70%	14.23%	15.24%
				30%	13.96%	13.59%	38.55%	13%	4.70%	13.28%	21.88%
				50%	13.96%	13.59%	38.55%				
25	14.74%	13.59%	39.94%	5%	8.37%	9.95%	28.58%	5%	8.60%	12.46%	32.37%
				15%	13.77%	13.59%	38.23%	8%	-0.31%	14.23%	9.30%
				30%	13.77%	13.59%	38.26%	13%	13.72%	13.59%	38.14%
				50%	13.77%	13.59%	38.26%				

[1] AR represents the accumulated return.

[2] MDD represents the maximum drawdown.

[3] SR represents the Sharpe ratio.

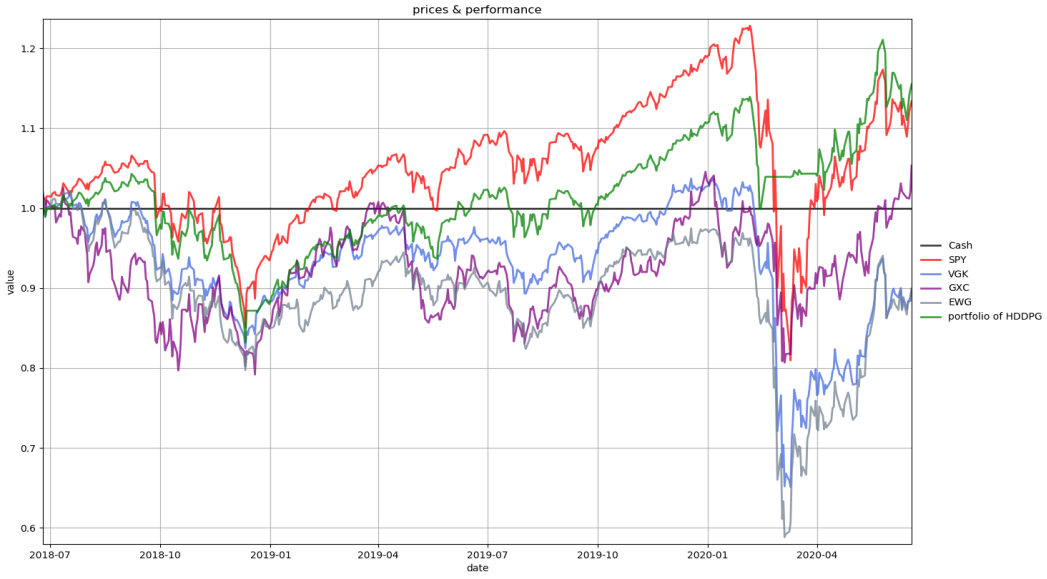


Figure 2.14: The price movements of each stock and portfolio values of Hierarchical DDPG with window size of ten-day and CVaR constraint $C = 5\%$.

the world's largest energy companies, which operates in integrated energy, chemicals, and petroleum operations in more than 180 countries worldwide. LUV typically refers to as Southwest Airlines Co., the world's largest low-cost airline offering cheaper air transportation in the United States. The data is collected from Yahoo Finance, consists of daily prices and volumes from 2010-01-01 to 2020-07-30, the same as the period of the ETF indexes. Then, the data is split into training and testing sets in the ratio of 8:2. The data preprocessing is implemented similarly as in Section 2.6.

As shown in Table 2.4, we obtain that the maximum drawdown of Distributional DDPG has significantly decreased for different window sizes compared to classical DDPG. We observe that the accumulated return and Sharpe ratio of Distributional DDPG tend to increase as α increases. Table 2.4 illustrates that Hierarchical DDPG outperforms classical DDPG in perspective of the maximum drawdown. Also, the Hierarchical DDPG algorithm can achieve a higher accumulated return and Sharpe ratio in many cases. For example, when the window size is 20 or 25, we can observe that the maximum drawdown of Hierarchical DDPG for both cases has significantly decreased and the Sharpe ratio has increased in most

cases in comparison with the classical DDPG method. We note that it is crucial to choose the appropriate window size and risk tolerance parameters for the superior performance of Hierarchical DDPG.

Table 2.4: Experimental Results for Additional Dataset

	DDPG			Distributional DDPG				Hierarchical DDPG			
Window size	AR	MDD	SR	α	AR	MDD	SR	C	AR	MDD	SR
5	7.15%	13.99%	26.44%	5%	0.00%	0.00%	0.00%	5%	12.99%	9.05%	41.39%
				15%	1.26%	7.81%	5.15%	8%	19.13%	9.05%	54.63%
				30%	6.19%	11.60%	24.86%	13%	21.88%	8.83%	59.45%
				50%	8.69%	12.97%	29.35%				
10	15.29%	15.73%	39.37%	5%	0.00%	0.00%	0.00%	5%	30.71%	8.26%	76.01%
				15%	1.19%	1.12%	25.05%	8%	12.13%	11.97%	38.44%
				30%	9.46%	13.62%	30.44%	13%	28.16%	14.16%	63.48%
				50%	27.18%	14.25%	56.97%				
20	16.90%	19.41%	30.17%	5%	0.00%	0.00%	0.00%	5%	27.27%	12.83%	66.68%
				15%	1.83%	8.66%	9.83%	8%	14.64%	12.97%	40.87%
				30%	21.85%	14.14%	49.62%	13%	20.11%	14.46%	48.82%
				50%	16.18%	16.34%	41.12%				
25	22.38%	16.34%	36.32%	5%	1.44%	0.41%	53.94%	5%	4.12%	10.69%	20.79%
				15%	5.61%	13.25%	23.58%	8%	35.33%	14.52%	66.98%
				30%	27.35%	14.16%	60.19%	13%	14.35%	16.13%	38.42%
				50%	37.53%	14.94%	63.38%				

[1] AR represents the accumulated return.

[2] MDD represents the maximum drawdown.

[3] SR represents the Sharpe ratio.

2.7 Conclusion

Portfolio management has always been a crucial topic in the financial field, which allocates investments in a group of assets to gain the maximum return of investors. It is a challenging task to construct a trading policy in the financial market because it requires professional knowledge in several fields, such as quantitative finance and risk management. The Deep RL algorithms can provide a more effective way to construct trading policies. Although Deep RL has achieved remarkable performance in portfolio management problems, most of the existing methods have not considered the worst-case scenarios in constructing trading policy. In this

chapter, we propose two novel approaches, Hierarchical DDPG and Distributional DDPG to address this issue.

To validate the applicability of the proposed learning analytics methods, a back-test is carried out on the real-world stocks from the U.S. financial market. Our study illustrates the superior performance of Hierarchical DDPG. It is impressive that the Hierarchical DDPG agent can not only maximize the portfolio profit but also keep the portfolio risk below a certain level of risk, which produces a portfolio with higher return and lower risk. Also, the experiment results reveal that Distributional DDPG produces risk-sensitive policies to reduce the effects of disaster events depending on the risk parameter. When the risk parameter α is small, the agent optimizes the performance for the worst-case scenario, which provides a conservative trading strategy. In contrast, when the risk parameter α is large, the agent is more willing to select an aggressive trading strategy. We can conclude that our Hierarchical DDPG and Distributional DDPG models outperform the classical DDPG method in the sense that they provide risk-sensitive strategies that protect investors who may suffer a huge loss caused by rare disaster events. Our proposed approaches provide effective methods to learn a risk-sensitive solution for the portfolio optimization problem.

The limitation of this work is that the distributional DDPG method is developed based on the assumption of the returns distribution that leads to a closed-form of calculation for the objective function. In addition, it is important to select the appropriate window size and risk tolerance parameters as needed for superior performance of the proposed models. For future research, we may involve textual data such as news or tweets, to improve the performance of DDPG, Hierarchical DDPG, and Distributional DDPG.

3 Robust optimal life insurance purchase and investment-consumption with regime-switching alpha-ambiguity maxmin utility

In this chapter, we delve into the intricate dynamics of life insurance intertwined with investment-consumption, a topic that has garnered significant attention in both actuarial science and financial economics. Through a rigorous exploration of models, theories, and practices, we aim to shed light on the multi-faceted influences of regime-switching and ambiguity aversion on these critical financial decisions.

3.1 Introduction

The optimal life insurance and investment-consumption problem is an important research topic in actuarial science and financial economics. Yarri (1965) establishes the life-cycle modelling framework and finds that without a bequest motive a rational investor should immediately convert all her savings into a life annuity, that is, she should fully annuitize the wealth, and consume all instantaneous annuity payments at every instant. The continuous-time version of consumption and investment problems is pioneered by Merton (1969), in which the stochastic optimal control theory is applied to develop an elegant solution to the problem. Richard (1975) extends the models of Merton (1969) and Yarri (1965) by introducing a life insurance purchase decision to the investment-consumption problem over a fixed planning horizon, but with an uncertain lifetime and a continuous lifetime income stream. An important finding in Richard (1975) is that the expected lifetime income has a positive effect on the demand for life insurance. Pliska and Ye (2007) study the optimal life insurance purchase and consumption/investment for a wage earner with a random and

unbounded lifetime with a given retirement time.

Traditionally, it is assumed that the dynamics of the asset price are governed by the geometric Brownian motion (GBM) model. However, the GBM model is unable to capture the long-term nature of the life-cycle model, whose horizon is as long as several decades. Over such a long horizon, the asset price is also affected by macroeconomic conditions and business cycles. Markovian regime-switching models stand out as a class of ideal candidates that can capture the macroeconomic regimes affecting the market environment, and allow the values of model parameters to change from one state to another over time. Therefore, regime-switching models provide a closer representation of reality compared to traditional models. The history of regime-switching models can be traced back to Hamilton (1989), who first develops the model for the stock return time series and demonstrates that regime-switching models could present the financial market more accurately than traditional models with deterministic coefficients. In addition, the regime-switching models can also provide a more efficient way to describe the effects of structural changes in different macroeconomic conditions. Zariphopoulou (1992) applies the regime-switching model to the investment-consumption problem, in which a financial market consists of a deterministic risk-free asset and a risky asset, depending only on a continuous-time Markov chain. Lee and Shim (2015) apply the Markovian regime-switching model to optimal consumption, investment, and life insurance purchase rules for a wage earner with mortality risk in a continuous time horizon, and apply the Markov chain approximation method to solve the Hamilton-Jacobi-Bellman (HJB) equation arising from the optimization problem.

Although the life insurance purchase problem has been studied extensively, little research has considered the effects of model uncertainty on dynamic life insurance decisions. Model uncertainty plays an essential role in the effectiveness of decision-making in insurance, finance, and other areas. In many circumstances, the individual is uncertain about a reference model, which may not accurately reflect the real situation. Therefore, any particular probability measure used to develop the model may be subject to a considerable degree of model misspecification. Model misspecification is usually caused by a lack of information regarding the probability measure, which is referred to as ambiguity. Research on model ambiguity is pioneered by Hansen and Sargent (2001), who consider the problem of asset pricing for

discrete time with model misspecification and propose the max-min expected utility theory to investigate model uncertainty. Anderson, Hansen, and Sargent (2003) extend the model in Hansen and Sargent (2001) to a continuous-time framework, by taking into account a set of alternative measures and quantifying the distance between the reference model and the alternative model via a relative entropy penalty term in the stochastic optimization procedure. Maenhout (2004) investigates the optimal investment-consumption selection problem with model uncertainty and obtains closed-form solutions by considering the homothetic robustness term in the dynamic investment and consumption problem. Elliott and Siu (2009) consider the robust optimal portfolio selection problem in a continuous-time Markov regime-switching market when an individual faces model uncertainty, and assume all the coefficients of the financial market are modulated by a two-state Markov chain, whose states are interpreted as the different states of an economy.

Most of the models proposed in the literature rely on the assumption that the individual admits the extremely ambiguity-averse attitude and aims to seek a robust optimal portfolio in the worst-case scenario. In a simplified setting, the robust optimal utility function is given by

$$\inf_{\mathbb{Q}\in\mathscr{Q}} \mathbb{E}^{\mathbb{Q}}\left[\int_0^T \big(U(c(t)) + \psi^{\mathbb{Q}}(t)\big)dt\right], \tag{3.1}$$

where U is the utility function, $\mathscr{Q}$ is a set of probability measures, $c(t)$ is the intertemporal consumption at time t, $\psi^{\mathbb{Q}}(t)$ is a function that penalizes the deviation of $\mathbb{Q}$ from the reference measure $\mathbb{P}$, and $\mathbb{E}^{\mathbb{Q}}$ is an expectation under $\mathbb{Q}$ measure. However, the robust utility model (3.1) is restrictive in that it only admits the extremely ambiguity-averse attitude; that is, it only considers the worst-case scenario. There are few behavioral experiments that support such an extreme pessimistic ambiguity attitude on the part of decision-makers. Heath and Tversky (1991) and Ghirardato, Maccheroni, and Marinacci (2004) conduct a series of experiments showing that people's reactions can be less ambiguity-averse or even ambiguity-seeking when they feel knowledgeable and experienced in some contexts. One can refer to Füllbrunn, Rau, and Weitzel (2014) and Dimmock, Kouwenberg, Mitchell, and Peijnenburg (2016) for further evidence. In light of complex attitudes towards ambiguity, Klibanoff, Marinacci, and Mukerji (2005, 2009) propose a more general utility form, namely the alpha-maxmin expected utility. This expected utility form allows the individual to have not only an aversion

attitude towards ambiguity but also a positive attitude towards ambiguity, i.e., ambiguity loving attitude.

As shown in human behavior experiments (Bier & Connell, 1994; Einhorn & Hogarth, 1985; Pulford, 2009), an optimistic individual has a low preference for ambiguity aversion; in contrast, a pessimistic individual has a higher preference for ambiguity aversion. On the other hand, the investor's sentiment changes over time. During an economic boom, the individual is more optimistic; on the contrary, in an economic recession, the individual is more pessimistic. Inspired by the above experimental evidence, in this chapter we assume that the individual has different levels of ambiguity aversion in different regimes, and we develop a novel utility function to capture the individual's regime-dependent ambiguity aversion. The utility function is integrated well with the alpha-maxmin expected utility (i.e., α-MEU) framework. Hence, throughout this chapter, this utility function is called a regime-switching alpha-ambiguity utility, which is a weighted sum of expected utility in the worst-case scenario and in the best-case scenario. The weights in regime-switching alpha-ambiguity utility are stochastic and depend on the regimes, which are modeled by a continuous-time Markov chain. The regime-switching alpha-ambiguity utility function allows the level of ambiguity aversion $\alpha(t)$ to change from one state to another. The main challenge of α-MEU framework is that the criterion is a linear aggregation of the expected utility under two distinct probability measures, which may cause dynamic inconsistency in decision-making. The precommitted strategy (Pirvu & Zhang, 2014; Zhou & Li, 2000) is one way to deal with dynamic inconsistency in optimal control problems in the literature. It is interpreted as "optimal from the point of view of time zero", and rational individuals follow the optimal strategy chosen at an initial time in the future. That is, the optimal strategy is derived under the assumption that the individuals precommit themselves not to deviate from the strategy chosen at the initial time and that is time-inconsistent. However, in many situations, time-consistency of strategy is a basic requirement for rational decision-makers, or today's preference may be different from tomorrow's preference. Beissner, Lin, and Riedel (2020) derive a dynamically consistent extension of the α-maxmin model for continuous time, and the time-consistent strategy retains the α-maxmin structure and allows distinction between ambiguity and ambiguity attitude. In this chapter, we consider the optimization problem when the objective

function changes over the time horizon and derive a corresponding optimal precommitted strategy under the proposed regime-switching alpha-ambiguity utility framework.

Under the regime-switching alpha-ambiguity utility framework we consider that the investor can dynamically purchase life insurance and allocate their wealth between the risky asset and the risk-free asset. She also receives an income at rate $\iota(t)$ continuously, which is terminated upon death or retirement, whichever happens first. We assume a simple financial market with regime-dependent market coefficients, which are modulated by an N-state continuous-time observable Markov chain. The life insurance has an instantaneous term: the higher the insurance premium rate the investor would like to pay, the more benefit her beneficiary will receive upon premature death. The investor aims to find robust optimal life insurance purchase and investment-consumption rules by maximizing the discounted version of the regime-switching alpha-ambiguity expected utility of intertemporal consumption, terminal wealth, and bequest if she dies before retirement.

The contribution of this chapter is threefold. First and foremost, the proposed regime-switching alpha-ambiguity utility integrates the regime-switching model with the alpha-maxmin utility framework seamlessly. That is, regime-switching models use a continuous-time Markov chain with finite-state space to represent the uncertainty of long-term macroeconomic factors, while the regime-switching alpha-maxmin expected utility inherits the idea of regime-switching models and allows individuals to have different attitudes towards ambiguity in different macroeconomic conditions. The robust optimal life insurance and investment-consumption strategies are solved under this new framework. Although model ambiguity has been well studied, the effect of an individual's ambiguity aversion on life insurance has been ignored. Thus, our second contribution is to pioneer a study on the effects of an individual's ambiguity aversion on the robust optimal life insurance decision. Moreover, most literature does not differentiate between ambiguity and ambiguity aversion. To differentiate between model ambiguity and the ambiguity aversion attitude, we borrow the scheme of ambiguity aversion and ambiguity settings from B. Li, Li, and Xiong (2016), in which mean-variance reinsurance-investment strategies are studied. Then, we derive the HJB equations by the stochastic dynamic programming principle, and the optimization problem is reduced to solving a system of HJB equations with new ingredients around model uncertainty and varying

attitudes toward uncertainty. Thirdly, we provide a novel verification theorem, which takes into account the best-case measure and the worst-case measure simultaneously and is tailor-made for the regime-switching alpha maxmin expected utility framework. This chapter also contains a number of numeric contributions. First, we find that ambiguity aversion has negative effects on optimal life insurance and consumption. A higher ambiguity aversion attitude makes the investor more concerned about the worst-case scenario, therefore prompting a more conservative strategy. Also, we find that the rational investor is more concerned about the model's uncertainty in the bear market, which is consistent with financial intuition. Second, our results highlight that the investor's expenditure on life insurance and consumption is higher in a bull market than it is in a bear market; and the effect of regime-switching on optimal life insurance and investment-consumption rules is strong when the investor is young, while the effect is negligible when it is close to her retirement age. On the other hand, we find that ambiguity aversion plays a pivotal role in utility loss if model uncertainty is absent. If the investor only has an extremely ambiguity-averse attitude, ignoring ambiguity may lead to drastic utility loss.

The remainder of this chapter is organized as follows. Section 3.2 introduces the market model dynamics and formulates the robust optimal life insurance and investment-consumption selection problem. Section 3.3 presents the Hamilton-Jacobi-Bellman (HJB) equation and verifies the existence and uniqueness of the solution to the HJB equation. Section 3.4 provides numerical examples and sensitivity analysis of robust optimal strategies and utility loss. Section 3.5 concludes this chapter. Technical proofs are provided in the appendices.

3.2 The Model Dynamics

In this section, we introduce a financial market and an insurance market that are available to the investor. The investment opportunities consist of one risky asset (i.e., stock) and one risk-free asset (i.e., bank account). Suppose that these two assets can be traded continuously in a finite-time horizon $\mathcal{T} := [0, T]$, where $T < \infty$. Market coefficients switch between N regimes, which are modulated by a finite-state Markov chain. We start by introducing the financial market, followed by the insurance market and the investor's mortality and wealth

processes. Finally, we formulate the investor's robust optimal selection problem under the regime-switching alpha maxmin utility model.

3.2.1 The Financial Market

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where $\mathbb{P}$ is a real-world probability measure. Let $B := \{B(t)|t \in \mathcal{T}\}$ be a one-dimensional standard Brownian motion, and $\mathcal{X} := \{\mathcal{X}(t)|t \in \mathcal{T}\}$ be an observable continuous-time, stationary, finite-state Markov chain on $(\Omega, \mathcal{F}, \mathbb{P})$. The state of the Markov chain $\mathcal{X}(t)$ corresponds to the state of an economy at time t. We identify the state space of the chain by the canonical state space, that is, $E := \{e_1, e_2, \ldots, e_N\}$, where $e_i \in \mathbb{R}^N$ denotes the i-th unit vector, of which the i-th component is one and the others are zero, for each $i = 1, 2, \ldots, N$. Throughout this chapter, we suppose that the Brownian motion B and the Markov chain $\mathcal{X}$ are stochastically independent. To specify the statistical properties of the Markov chain $\mathcal{X}$, we define a transition matrix $\mathcal{Q} := [q_{ij}]_{N\times N}$, where q_{ij} is the instantaneous transition rate of the Markov chain from state j to state i. Note that for each $i, j = 1, 2, \ldots, N$, $q_{ij} > 0$, for $i \neq j$ and $\sum_{j=1}^{N} q_{ij} = 0$, so $q_{ii} < 0$. Let $\mathbb{F}^{\mathcal{X}} := \{\mathcal{F}^{\mathcal{X}}(t)|t \in \mathcal{T}\}$ be the right-continuous, $\mathbb{P}$-complete filtration generated by the Markov chain $\mathcal{X}$. According to Elliott, Aggoun, and Moore (1995), we decompose the chain $\mathcal{X}$ using the following semi-martingale dynamics:

$$\mathcal{X}(t) = \mathcal{X}(0) + \int_0^t \mathcal{Q}\mathcal{X}(s)ds + M(t), \quad t \in \mathcal{T}.$$

Here, $M := \{M(t)|t \in \mathcal{T}\}$ is an $\mathbb{R}^N$-valued, $(\mathbb{F}^{\mathcal{X}}, \mathbb{P})$-martingale.

In what follows, we introduce the model dynamics of the primitive assets. Let r be the risk-free interest rate, which is assumed to be a positive constant. The price process of the risk-free asset $P(t)$ is governed by

$$dP(t) = rP(t)dt, \quad P(0) = 1.$$

Let $\mu(t)$ and $\sigma(t)$ be the expected return rate and the volatility of the risky asset price at time t. We assume that they are modulated by the Markov chain $\mathcal{X}$ as follows:

$$\mu(t) := \langle \boldsymbol{\mu}, \mathcal{X}(t) \rangle \qquad \text{and} \qquad \sigma(t) := \langle \boldsymbol{\sigma}, \mathcal{X}(t) \rangle,$$

where $\boldsymbol{\mu} := (\mu_1, \mu_2, \ldots, \mu_N)' \in \mathbb{R}^N$ with $\mu_i > r$ and $\boldsymbol{\sigma} := (\sigma_1, \sigma_2, \ldots, \sigma_N)' \in \mathbb{R}^N$ with $\sigma_i > 0$; μ_i and σ_i are the expected return rate and volatility of the risky asset when the economy is in the i-th state, respectively, for each $i = 1, 2, ..., N$. The risky asset price process $S(t)$ follows a Markov-modulated Geometric Brownian Motion (GBM) model:

$$dS(t) = \mu(t)S(t)dt + \sigma(t)S(t)dB(t), \quad S(0) = S_0 > 0.$$

Now, we are ready to describe the information structure. Let $\mathbb{F}^B := \{\mathcal{F}^B(t)|t \in \mathcal{T}\}$ be the right-continuous, $\mathbb{P}$-complete filtration generated by the Brownian motion, $\mathbb{F} := \{\mathcal{F}(t)|t \in \mathcal{T}\}$ be the enlarged filtration generated by the Brownian motion and the chain, and $\mathcal{F}(t) := \mathcal{F}^{\mathcal{X}}(t) \vee \mathcal{F}^B(t)$ be the enlarged σ-field of $\mathcal{F}^{\mathcal{X}}(t)$ and $\mathcal{F}^B(t)$.

3.2.2 The Life Insurance Market

We assume that the investor is alive at time $t = 0$ and has a remaining lifetime τ, which is a nonnegative random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that the random variable τ has probability density function $f(t)$, distribution function $F(t)$, and survival function $\bar{F}(t)$ such that

$$F(t) = P(\tau < t) = \int_0^t f(u)du \qquad \text{and} \qquad \bar{F}(t) = P(\tau \geq t) = 1 - F(t).$$

The hazard function $\lambda(t)$ represents the instantaneous death rate for the investor at current age y who has survived to time t, and it is defined by

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t | \tau \geq t)}{\Delta t}.$$

Here, $\lambda(t)$ is assumed to be a given deterministic Borel measureable function such that $\int_0^\infty \lambda(t)dt = \infty$.

Let $f(s,t)$ and $\bar{F}(s,t)$ denote the conditional probability density for death and the conditional probability of being alive, respectively, at time s, conditional upon being alive at time $t \leq s$. As shown in Lee and Shim (2015); Pliska and Ye (2007), we have

$$f(s,t) = \lambda(s)e^{-\int_t^s \lambda(v)dv} \qquad \text{and} \qquad \bar{F}(s,t) = e^{-\int_t^s \lambda(v)dv}.$$

In the life insurance market, the investor purchases life insurance to protect her family in case of premature death. We assume that the life insurance is offered continuously, and the

investor chooses a life insurance contract by paying premiums at the rate $p(t)$ to the insurance company. In compensation, if the investor dies at time $t < T$, then the insurance company pays an insurance benefit at the amount of $p(t)/\eta(t)$, where insurance premium-payout ratio $\eta(t)$ is a continuous and deterministic function determined by the insurance company. The contract will be terminated if the investor dies before time T or achieves retirement at time T. Therefore, the investor's total bequest to her beneficiary in the event of death at $t < T$ is given by

$$Z(t) = W(t) + \frac{p(t)}{\eta(t)},$$

where $W(t)$ denotes the investor's financial wealth at time t.

3.2.3 The Wealth Process

Suppose that the investor is endowed with the initial wealth W_0 and will receive an income rate $\iota(t)$ till $\min\{T, \tau\}$. In other words, the income will be terminated by the investor's death or retirement at time T, whichever happens first. Furthermore, we assume the function $\iota(t)$ is a Borel measurable function satisfying the integrability condition

$$\int_0^T \iota(s)ds < \infty.$$

We denote by $W := \{W(t)|t \in \mathcal{T}\}$ the investor's wealth process. Let $\pi(t)$ be the proportion of the wealth invested in the risky asset, and $c(t)$ be the amount of the wealth for consumption. In this chapter, we assume that the short-selling is allowed in the market, that is, π can take negative values, and there are no transaction costs and taxes. Therefore, with a triplet of investment, consumption and insurance strategies $(\pi(t), c(t), p(t))$, the investor's wealth process $W(t)$ is governed by the following stochastic differential equation:

$$dW(t) = \Big\{ \big[\mu(t) - r\big]\pi(t)W(t) + rW(t) - c(t) - p(t) + \iota(t) \Big\} dt + \pi(t)\sigma(t)W(t)dB(t) \quad (3.2)$$

with initial wealth $W(0) = W_0 > 0$ and initial state $\mathcal{X}(0) = e_i \in E$.

3.2.4 Regime-switching alpha-ambiguity maxmin utility

In this subsection, we first introduce a novel regime-switching alpha-ambiguity maxmin expected utility, which describes the individual's attitude towards risk and ambiguity. We

then formulate the robust optimal life insurance and investment-consumption selection rules for the investor, who aims at maximizing the regime-switching alpha-ambiguity expected discounted utility of her intertemporal consumption, bequest if she dies before retirement time T, and terminal wealth. We assume that the investor's preference is measured by the Constant Relative Risk Aversion (CRRA) utility function as follows:

$$U(v) = \frac{v^{1-\gamma}}{1-\gamma}, \quad v > 0,$$

where γ is the risk aversion parameter such that $\gamma > 0$ and $\gamma \neq 1$.

As pointed out by Anderson et al. (2003), the individual accepts the reference model as useful, but suspects it is misspecified. Therefore, the individual seeks a pool of alternative models and considers the penalty that may occur if the alternative models deviate too far away from the reference model. Inspired by Maenhout (2004), we employ a relative entropy to measure the "distance" between the reference model and the alternative models. Indeed, the relative entropy is also considered as a risk measurement in Siu (2011) under regime-switching models. Let $\theta := \{\theta(t)|t \in \mathcal{T}\}$ and $\phi_{ij} := \{\phi_{ij}(t)|t \in \mathcal{T}\}$ be predictable processes in $\mathbb{R}$ and $\mathbb{R}^+$, respectively. To simplify the following presentation, we define a set $\phi := \{\phi_{ij}|i,j = 1,2,\ldots,N, i \neq j\} \subset (\mathbb{R}^+)^\ell$, where $\ell := N \times N - N$. Thus, an alternative probability measure $\mathbb{Q}$ is defined by the Radon–Nikodym derivative as below:

$$\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_T} = \Lambda(T) := \Lambda_{BM}(T) \times \Lambda_{MC}(T), \tag{3.3}$$

where

$$\Lambda_{BM}(T) := \exp\left\{\int_0^T \theta(t)dB(t) - \frac{1}{2}\int_0^T (\theta(t))^2 dt\right\}, \tag{3.4}$$

and

$$\begin{aligned}\Lambda_{MC}(T) := \exp\Bigg\{ & \sum_{i,j=1,i\neq j}^{N} \int_0^T \log\phi_{ij}(t)dM_{ij}(t) \\ & + \sum_{i,j=1,i\neq j}^{N} \int_0^T \Big[\phi_{ij}(t)\log\phi_{ij}(t) - \phi_{ij}(t) + 1\Big] q_{ij}\mathbb{1}_{\{\mathcal{X}(t^-)=e_i\}}dt\Bigg\}.\end{aligned} \tag{3.5}$$

Here, $M_{ij}(t) := \int_0^t \langle \mathcal{X}(s^-), e_i\rangle\langle dM(s), e_j\rangle$ is an $(\mathbb{F}^{\mathcal{X}}, \mathbb{P})$-martingale, for each $i,j = 1,2,...,N$, and $i \neq j$.

In the above Radon–Nikodym derivative, the first equation (3.4) changes the dynamics of the stock price by adjusting the expected stock return, and the second equation (3.5) changes the probability law of the Markov chain. Precisely, under the alternative measure $\mathbb{Q}$, it follows from Girsanov's theorem that

$$B^{\mathbb{Q}}(t) := B(t) - \int_0^t \theta(s)ds,$$

is a standard Brownian motion; the intensity of the Markov chain is given by:

$$q_{ij}^{\phi}(t) := \phi_{ij}(t)q_{ij}, \quad \text{where} \quad i \neq j, \qquad \text{and} \qquad q_{ii}^{\phi}(t) := -\sum_{j=1, j\neq i}^{N} \phi_{ij}(t)q_{ij}.$$

Under the $\mathbb{Q}$ measure, we define a new transition matrix $\mathcal{Q}^{\phi}(t) := [q_{ij}^{\phi}(t)]_{N\times N}$. Thus, the semi-martingale decomposition of chain $\mathcal{X}$ has the following representation under the $\mathbb{Q}$ measure:

$$\mathcal{X}(t) = \mathcal{X}(0) + \int_0^t \mathcal{Q}^{\phi}(s)\mathcal{X}(s)ds + M^{\phi}(t), \quad t \in [0, T].$$

Here, $M^{\phi}(t)$ is an $\mathbb{R}^N$-valued, $(\mathbb{F}^{\mathcal{X}}, \mathbb{Q})$-martingale. Denote by

$$M_{ij}^{\phi}(t) := \int_0^t \langle \mathcal{X}(s^-), e_i\rangle \langle dM^{\phi}(s), e_j\rangle,$$

an $(\mathbb{F}^{\mathcal{X}}, \mathbb{Q})$-martingale, for each $i, j = 1, 2, ..., N$, and $i \neq j$.

Therefore, under the $\mathbb{Q}$ measure, the risky asset price is governed by:

$$dS^{\mathbb{Q}}(t) = S(t)\big\{[\mu(t) + \sigma(t)\theta(t)]dt + \sigma(t)dB^{\mathbb{Q}}(t)\big\},$$

and the wealth process evolves:

$$\begin{aligned} dW^{\mathbb{Q}}(t) =& \Big\{\big[\mu(t) - r\big]\pi(t)W(t) + rW(t) - c(t) - p(t) + \iota(t)\Big\}dt + \theta(t)\pi(t)\sigma(t)W(t)dt \\ &+ \pi(t)\sigma(t)W(t)dB^{\mathbb{Q}}(t). \end{aligned} \tag{3.6}$$

The relative entropy is defined as the $\mathbb{Q}$-expectation of the log Radon-Nikodym derivative. Then, it follows from equation (3.3) that the relative entropy over the interval $[0, T]$ is given

by:

$$\begin{aligned}\mathcal{D}(\mathbb{Q},\mathbb{P}) &:= \mathbb{E}^{\mathbb{Q}}\big[\log\Lambda(T)\big]\\ &= \mathbb{E}^{\mathbb{Q}}\Bigg[\int_0^T \theta(t)dB(t) - \frac{1}{2}\int_0^T(\theta(t))^2dt + \sum_{i,j=1,i\neq j}^{N}\int_0^T \log\phi_{ij}(t)dM_{ij}(t)\\ &\quad + \sum_{i,j=1,i\neq j}^{N}\int_0^T\Big[\log\phi_{ij}(t) - \phi_{ij}(t) + 1\Big]q_{ij}\mathbb{1}_{\{\mathcal{X}(t^-)=e_i\}}dt\Bigg]\\ &= \mathbb{E}^{\mathbb{Q}}\Bigg[\frac{1}{2}\int_0^T(\theta(t))^2dt + \sum_{i,j=1,i\neq j}^{N}\int_0^T\Big[\phi_{ij}(t)\log\phi_{ij}(t) - \phi_{ij}(t) + 1\Big]q_{ij}\mathbb{1}_{\{\mathcal{X}(t^-)=e_i\}}dt\Bigg].\end{aligned}$$

Here, $\mathbb{E}^{\mathbb{Q}}$ is an expectation under the $\mathbb{Q}$ measure.

Thus, we define the "penalty" generator between the alternative and reference models by the normalized relative entropy:

$$\psi(t) = \frac{\theta(t)^2}{2\Psi(t,W,e_i)} + \frac{\sum_{i,j=1,i\neq j}^{N} q_{ij}\Big[\phi_{ij}(t)\log\phi_{ij}(t) - \phi_{ij}(t) + 1\Big]}{\Psi(t,W,e_i)}, \tag{3.7}$$

where $\Psi(t,W,e_i)$ reflects the preference levels of robustness. Denoting by $\beta(t)$ the level of robustness at time t, we assume that the $\beta(t)$ is modulated by the Markov chain as follows:

$$\beta(t) := \langle\boldsymbol{\beta},\mathcal{X}(t)\rangle,$$

where $\boldsymbol{\beta} := (\beta_1,\beta_2,\ldots,\beta_N)' \in \mathbb{R}^N$ with $\beta_i > 0$.

Throughout this chapter, we adopt the parametric form suggested by Maenhout (2004):

$$\Psi(t,W,e_i) = \frac{\beta(t)}{(1-\gamma)V(t,W,e_i)}, \quad \beta(t) > 0. \tag{3.8}$$

This parametric form (3.8) is economically meaningful and facilitates analytical tractability. Intuitively, when selecting adverse drift distortions $\theta(t)$ and adverse transition rate distortions $\phi_{ij}(t)$ in equation (3.7), and moving away from the reference model, the incurred entropy penalties are weighted by $\frac{1}{\Psi(t,W,e_i)}$. Therefore, if the investor has less faith in the reference model, then she will be eager to make more robust decisions.

Now, we are ready to introduce the set of admissible strategies, which is defined as follows.

Definition 3.2.1 *A triplet* $\{(\pi(t),c(t),p(t))|t \in \mathcal{T}\}$ *is called an admissible investment-consumption-insurance strategy if*

(i) π, c, *and* p *are progressively measurable processes in* $\mathbb{R}$, $\mathbb{R}^+$, *and* $\mathbb{R}$, *respectively;*

(ii) π, c, *and* p *are almost surely integrable in the sense such that*

$$\int_0^T (|\pi(t)|^2 + |c(t)| + |p(t)|)dt < \infty, \quad \mathbb{P} - a.s.;$$

(iii) the stochastic differential equation (3.2) *associated with* (π, c, p) *has a unique strong solution* $\{W(t)|t \in \mathcal{T}\}$.

The space of all admissible strategies is denoted by $\mathcal{A}$.

The set of $\{(\theta(t), \phi(t))|t \in \mathcal{T}]\}$ *is called an admissible set of distortion processes such that*

(i) $\theta(t)$ *and* $\phi_{ij}(t)$ *are predictable processes in* $\mathbb{R}$ *and* $\mathbb{R}^+$, *respectively, for each* $t \in \mathcal{T}$;

(ii) $\mathbb{E}^{\mathbb{Q}}\left[\exp\left\{\frac{1}{2}\int_0^T(\theta(t))^2dt + \sum_{i,j=1,i\neq j}^{N}\int_0^T\left[\phi_{ij}(t)\log\phi_{ij}(t) - \phi_{ij}(t) + 1\right]q_{ij}\mathbb{1}_{\{\mathcal{X}(t^-)=e_i\}}dt\right\}\right] < \infty$.

The space of all admissible sets is denoted by Θ.

As discussed in the introduction, in the classical literature on robust decision-making, the individual is assumed to have an extremely ambiguity-averse attitude and aims to seek a robust investment in the worst-case scenario. However, Heath and Tversky (1991) and Klibanoff et al. (2005, 2009) provide a theoretical background, demystifying that individuals may be less ambiguity-averse or even ambiguity-seeking when they feel knowledgeable, experienced, or competent in the relevant context. Inspired by this observation, we consider a general case in which the wage earner has different attitudes towards ambiguity in different regimes.

Denote the subjective discount rate at time t by $\delta(t)$, which is determined by the Markov chain as follows:

$$\delta(t) := \langle \boldsymbol{\delta}, \mathcal{X}(t) \rangle,$$

where $\boldsymbol{\delta} := (\delta_1, \delta_2, \ldots, \delta_N)' \in \mathbb{R}^N$ with $\delta_i > 0$. Modeling the investor's time preference by the Markov-modulated discount rate reflects that their time preference is affected by economic factors.

Let $\alpha(t)$ denote the level of ambiguity aversion at time t, which is modulated by the Markov chain as follows:

$$\alpha(t) := \langle \boldsymbol{\alpha}, \mathcal{X}(t) \rangle,$$

where

$$\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_N)' \in \mathbb{R}^N.$$

Inspired by B. Li et al. (2016), we propose a new utility form, namely the regime-switching alpha-ambiguity expected utility, given by

$$\begin{aligned}
&\inf_{\mathbb{Q}\in\mathscr{Q}} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\bigg[\int_t^T \alpha(s) e^{-\int_t^s \delta(v)dv}\big(U\big(c(s)\big) + \psi^{\mathbb{Q}}(s)\big)ds\bigg] \\
&+ \sup_{\mathbb{Q}\in\mathscr{Q}} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\bigg[\int_t^T \hat{\alpha}(s) e^{-\int_t^s \delta(v)dv}\big(U\big(c(s)\big) - \psi^{\mathbb{Q}}(s)\big)ds\bigg],
\end{aligned}$$

where $\hat{\alpha}(s) := 1 - \alpha(s)$, and the first (resp. second) expectation corresponds to the individual's ambiguity-averse (resp. ambiguity-seeking) attitude. Here, $\mathscr{Q}$ is a set of probability measures. Throughout this chapter, we denote $\mathbb{E}^{\mathbb{Q}}_{t,w,i}[\cdot] = \mathbb{E}^{\mathbb{Q}}[\cdot|W(t) = w, \mathcal{X}(t) = e_i]$, the conditional expectation under $\mathbb{Q}$ given $W(t) = w$ and $\mathcal{X}(t) = e_i$. In particular, $\alpha(t) = 1, \frac{1}{2}, 0$ corresponds to an extremely ambiguity-averse, ambiguity-neutral, and extremely ambiguity-seeking attitude.

Throughout this chapter, we impose the following restriction on the range of $\alpha(t)$:

$$\frac{1}{2} \leq \alpha(t) \leq 1.$$

The above assumption is mainly due to the following two reasons. First, according to human behavior experiments (Ghirardato et al., 2004; Heath & Tversky, 1991; Maccheroni, Marinacci, & Rustichini, 2006; Maccheroni et al., 2006), most people not only have an ambiguity-seeking attitude in the best-case scenario, they also have an ambiguity-averse attitude in the best-case scenario. Second, many psychological experiments and theories, e.g., Tversky and Kahneman (1991) and Kahneman, Knetsch, and Thaler (1991) show that most people are more concerned about the worst-case scenario than the best-case scenario because people tend to avoid losses when they face the same amount of gains and losses.

One of the key features of regime-switching alpha-ambiguity expected utility is that it allows different levels of ambiguity aversion in different scenarios and states corresponding

to different macroeconomic circumstances, specified by $\alpha(t)$, and different levels of model ambiguity in different states, specified by $\beta(t)$. The advantage of this utility is that not only market coefficients but also the ambiguity aversion coefficients can switch from one state to another. Moreover, this novel utility takes into account the worst-case scenario and the best-case scenario simultaneously. When the economy is booming, the individual is optimistic about the future and cares more about the best-case scenario of economic prosperity. Hence, the optimal solution for the best-case scenario corresponds to a maximal expected utility, which leads to a "max-max" optimization problem. In contrast, when the economy is shrinking, the individual is pessimistic about the future and is more concerned about the worst-case scenario of economic recession. Thus, the optimal strategy for the worst-case scenario corresponds to a minimal expected utility, which leads to the "max-min" of expected utility. In general, the individual is more (resp. less) ambiguity-averse in economic recession (resp. economic prosperity). For example, in a two-state case (i.e., $N = 2$) with a bear market in state e_1 and a bull market in state e_2, respectively. Then the investor has a stronger ambiguity-averse attitude in the bear market than in the bull market. That is,

$$\alpha_1 \geq \alpha_2 \quad \text{and} \quad \hat{\alpha}_1 \leq \hat{\alpha}_2.$$

Motivated by the seminal work of Pliska and Ye (2007), we formulate the robust optimal life insurance purchase and investment-consumption problem as:

$$\begin{aligned}
\sup_{(\pi,c,p)\in\mathcal{A}} \Bigg\{ & \inf_{(\theta,\phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^{T\wedge\tau} e^{-\int_t^s \delta(v)dv}\big(\alpha(s)U\big(c(s)\big) + \psi(s)\big)ds \\
& + \alpha(\tau)e^{-\int_t^\tau \delta(v)dv}\xi(\tau)U\big(Z(\tau)\big)\mathbb{1}_{\{\tau\leq T\}} + \alpha(T)e^{-\int_t^T \delta(v)dv}\zeta(T)U\big(W(T)\big)\mathbb{1}_{\{\tau>T\}}\Bigg] \\
& + \sup_{(\theta,\Phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^{T\wedge\tau} \hat{\alpha}(t)e^{-\int_t^s \delta(v)dv}\big(U\big(c(s)\big) - \psi(s)\big)ds \\
& + \hat{\alpha}(\tau)e^{-\int_t^\tau \delta(v)dv}\xi(\tau)U\big(Z(\tau)\big)\mathbb{1}_{\{\tau\leq T\}} + \hat{\alpha}(T)e^{-\int_t^T \delta(v)dv}\zeta(T)U\big(W(T)\big)\mathbb{1}_{\{\tau>T\}}\Bigg]\Bigg\},
\end{aligned} \tag{3.9}$$

where $T \wedge \tau = \min[T, \tau]$. Here $\xi(t)$ and $\zeta(t)$ denote the weights of utility of bequest and terminal wealth at time t, which are modulated by the Markov chain as follows:

$$\xi(t) := \langle \boldsymbol{\xi}, \mathcal{X}(t)\rangle \qquad \text{and} \qquad \zeta(t) := \langle \boldsymbol{\zeta}, \mathcal{X}(t)\rangle,$$

where $\boldsymbol{\xi} := (\xi_1, \xi_2, \ldots, \xi_N)' \in \mathbb{R}^N$ and $\boldsymbol{\zeta} := (\zeta_1, \zeta_2, \ldots, \zeta_N)' \in \mathbb{R}^N$.

The next step is to transform problem (3.9) with the random planning horizon to an equivalent problem with a fixed planning horizon. Since τ is independent of the filtration $\mathcal{F}$, (3.9) is equivalent to:

$$\begin{aligned} V(t,w,e_i) = \sup_{(\pi,c,p)\in\mathcal{A}} \Bigg\{ \inf_{(\theta,\phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i} \Bigg[\int_t^T \alpha(s) e^{-\int_t^s \delta(v)dv} \Big[\bar{F}(s,t) U\big(c(s)\big) + \bar{F}(s,t)\psi(s) \\ + f(s,t)\xi(s) U\big(Z(s)\big) \Big] ds + \alpha(T) e^{-\int_t^T \delta(v)dv} \bar{F}(T,t)\zeta(T) U\big(W(T)\big) \Bigg] \\ + \sup_{(\theta,\phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i} \Bigg[\int_t^T \hat{\alpha}(s) e^{-\int_t^s \delta(v)dv} \Big[\bar{F}(s,t) U\big(c(s)\big) - \bar{F}(s,t)\psi(s) \\ + f(s,t)\xi(s) U\big(Z(s)\big) \Big] ds + \hat{\alpha}(T) e^{-\int_t^T \delta(v)dv} \bar{F}(T,t)\zeta(T) U\big(W(T)\big) \Bigg] \Bigg\}, \end{aligned} \tag{3.10}$$

where $V(t,w,e_i)$ is the value function of the problem.

Let us remark that the robust optimal life insurance and investment-consumption decisions can be obtained by maximizing the weighted sum of regime-switching alpha-ambiguity expected discounted utility of consumption, bequest if the investor dies before retirement, and terminal wealth. Specifically, the "infimum" part can be interpreted as the expected discounted utility of consumption, bequest, and terminal wealth in the worst-case scenario; the "supremum" part represents those in the best-case scenario.

3.3 Main Results

In this section, by the Hamilton-Jacobi-Bellman (HJB) equation approach, we derive the optimal solution for the robust optimal life insurance purchase and investment-consumption problem (3.10).

For that purpose, we define the differential generator $\mathcal{L}^{(\pi,c,p;\theta,\Phi)}$ on $V(t,w,e_i) \in C^{1,2}(\mathcal{T} \times \mathbb{R})$, for each $e_i \in E$, as follows:

$$\begin{aligned} \mathcal{L}^{(\pi,c,p;\theta,\phi)} V(t,w,e_i) :=& V_t + \frac{1}{2}\pi^2\sigma_i^2 w^2 V_{ww} + \big[(\mu_i - r)\pi w + rw - c - p + \iota(t) + \theta\pi\sigma_i w\big] V_w \\ & - \delta_i V - \lambda(t) V + \langle \mathcal{Q}^{\phi} e_i, \mathbf{V}(t,w) \rangle, \end{aligned} \tag{3.11}$$

where $\mathbf{V}(t,w) := \big(V(t,w,e_1), V(t,w,e_2), \cdots, V(t,w,e_N)\big)'$ is an N-dimensional vector, V_t, V_w, and V_{ww} represent partial derivatives $\frac{\partial V}{\partial t}$, $\frac{\partial V}{\partial w}$, and $\frac{\partial^2 V}{\partial w^2}$, respectively. Here $C^{1,2}(\mathcal{T}\times\mathbb{R})$ denotes a class of functions that are continuously differentiable with respect to t and twice continuously differentiable with respect to w on $\mathbb{R}$.

According to the dynamic programming principle, we can solve the following regime-switching HJB equation to find the solution to problem (3.10):

$$\begin{aligned}&\sup_{(\pi,c,p)\in\mathcal{A}}\left\{\inf_{(\theta,\phi)\in\Theta}\left\{\alpha_i\Big[\mathcal{L}^{(\pi,c,p;\theta,\phi)}V(t,w,e_i)+U(c)+\psi(t,w,e_i)+\lambda(t)\xi_i U\Big(w+\frac{p}{\eta(t)}\Big)\Big]\right\}\right.\\&\left.+\sup_{(\theta,\phi)\in\Theta}\left\{\hat{\alpha}_i\Big[\mathcal{L}^{(\pi,c,p;\theta,\phi)}V(t,w,e_i)+U(c)-\psi(t,w,e_i)+\lambda(t)\xi_i U\Big(w+\frac{p}{\eta(t)}\Big)\Big]\right\}\right\}=0\end{aligned}\tag{3.12}$$

with boundary condition $V(T,w,e_i)=\zeta_i U(w)$ for each $e_i\in E$ and $i=1,2,\ldots,N$.

3.3.1 Robust Optimal Solutions

In this subsection, by solving the HJB equation (3.12), we derive candidate optimal strategies to the robust investment-consumption-insurance problem (3.10) in semi-closed form.

Theorem 3.3.1 *For the robust optimal life insurance purchase and investment-consumption problem (3.10), when $\mathcal{X}(t)=e_i$, the value function is given by*

$$V(t,w,e_i)=f(t,e_i)\frac{[w+g(t)]^{1-\gamma}}{1-\gamma}.\tag{3.13}$$

the candidate robust optimal strategy $u^*(t):=(\pi^*(t),c^*(t),p^*(t))$ is given by

$$\begin{cases}\pi^*(t)=\frac{\mu_i-r}{\sigma_i^2}\frac{w+g(t)}{w\big[\gamma-(\hat{\alpha}_i-\alpha_i)\beta_i\big]},\\ c^*(t)=f^{-\frac{1}{\gamma}}(t,e_i)\big[w+g(t)\big],\\ p^*(t)=[\xi_i\lambda(t)]^{\frac{1}{\gamma}}f^{-\frac{1}{\gamma}}(t,e_i)\big[w+g(t)\big]\eta^{1-\frac{1}{\gamma}}(t)-w\eta(t),\end{cases}\tag{3.14}$$

the candidate worst-case measure is determined by

$$\begin{cases}\underline{\theta}^*(t)=-\frac{\mu_i-r}{\sigma_i}\frac{\beta_i}{\gamma-(\hat{\alpha}_i-\alpha_i)\beta_i},\\ \underline{\phi}_{ij}^*(t)=\exp\left\{\frac{\beta_i}{1-\gamma}\big[1-\frac{f(t,e_j)}{f(t,e_i)}\big]\right\},\end{cases}\tag{3.15}$$

and the candidate best-case measure is determined by

$$\begin{cases} \overline{\theta}^*(t) = \frac{\mu_i - r}{\sigma_i} \frac{\beta_i}{\gamma - (\hat{\alpha}_i - \alpha_i)\beta_i}, \\ \overline{\phi_{ij}^*}(t) = \exp\left\{ -\frac{\beta_i}{1-\gamma} \left[1 - \frac{f(t,e_j)}{f(t,e_i)} \right] \right\}, \end{cases} \tag{3.16}$$

where $f(t,e_i)$ and $g(t)$ satisfy the following equations:

$$\begin{aligned} 0 =& \frac{df(t,e_i)}{dt} + \Big[1 + [\xi_i \lambda(t)]^{\frac{1}{\gamma}} \eta^{1-\frac{1}{\gamma}}(t) \Big] \gamma f^{1-\frac{1}{\gamma}}(t,e_i) + \frac{1-\gamma}{\beta_i} \Big[\alpha_i \sum_{j=1, j\neq i}^{N} q_{ij} \Big(\underline{\phi_{ij}^*}(t) \log \underline{\phi_{ij}^*}(t) - \underline{\phi_{ij}^*}(t) + 1 \Big) \\ & - \hat{\alpha}_i \sum_{j=1, j\neq i}^{N} q_{ij} \Big(\overline{\phi_{ij}^*}(t) \log \overline{\phi_{ij}^*}(t) - \overline{\phi_{ij}^*}(t) + 1 \Big) \Big] f(t,e_i) + \sum_{j=1}^{N} q_{ij} f(t,e_j) \Big[\alpha_i \underline{\phi_{ij}^*}(t) + \hat{\alpha}_i \overline{\phi_{ij}^*}(t) \Big] \\ & - b_i(t) f(t,e_i), \end{aligned} \tag{3.17}$$

and

$$g(t) = \int_t^T \iota(s) \exp\left(- \int_t^s r + \eta(v) dv \right) ds, \tag{3.18}$$

with

$$b_i(t) := \delta_i + \lambda(t) + \frac{1}{2} \frac{(\mu_i - r)^2}{\sigma_i^2} \frac{1-\gamma}{[\hat{\alpha}_i - \alpha_i]\beta_i - \gamma} - r(1-\gamma) - \eta(t)(1-\gamma),$$

and terminal conditions $f(T,e_i) = \zeta_i$ and $g(T) = 0$.

Proof. See Appendix B.1. ■

Remark 3.3.1 *The function $g(t)$ can be considered as the human capital, that is, the actuarial present value of the wage earner's future income from time t to T, while $w + g(t)$ is the total wealth of the wage earner at time t, consisting of the current wealth and future income. According to equation (3.14), the optimal investment portfolio rule $\pi(t)$ also depends on the financial market regime, which is proportional to the Metron ratio $\frac{\mu(t)-r}{\sigma^2(t)}$, and is revised by total wealth $w + g(t)$ and ambiguity aversion $\alpha(t)$. The optimal consumption $c(t)$ in (3.14) can be thought as a proportion of the total wealth, and is weighted by the function of $f^{-\frac{1}{\gamma}}(t,e_i)$. The optimal life insurance $p(t)$ in (3.14) can be considered as a proportion of the total wealth, and is adjusted by the function of $f^{-\frac{1}{\gamma}}(t,e_i)$, mortality rate $\lambda(t)$, and insurance premium-payout ratio η. Moreover, $Z^*(t) = \frac{p^*(t)}{\eta(t)} + w$ is the optimal amount of legacy in the event of death at time t, which is proportional to the wage earner's total wealth $w + g(t)$.*

Remark 3.3.2 *Note that the equation (3.17) is a system of N-coupled, non-linear ODEs. In general, it is difficult to find a closed-form solution to this equation. Inspired by Theorem 3.1 in Shen and Siu (2013), we prove the existence and uniqueness of a solution $f(t, e_i)$ to (3.17). The following theorem summarizes the existence and uniqueness result.*

Theorem 3.3.2 *The non-linear ODE system (3.17) has a unique solution $f(t, e_i)$ for each $e_i \in E$ that is bounded and strictly positive over time t, for any $t \in \mathcal{T}$.*

Proof. See Appendix B.2. ■

3.3.2 Verification

Having solved the HJB equation (3.12), we are in a position to verify that the obtained candidates are true optimal strategies of the problem (3.10). The following theorem gives us the conditions under which the solution of the HJB equation is indeed the value function and the candidates are the optimal strategies.

Theorem 3.3.3 *For the robust optimal life insurance and investment-consumption problem (3.10), suppose that the function $V(t, w, e_i)$ is a classical solution to the HJB equation (3.12) such that the following conditions are satisfied:*

1. $\big\{V(t, W(t), \mathcal{X}(t))\big| t \in [0, T]\big\}$ *is uniformly integrable under both the worst-case and best-case scenarios;*

2. $\big\{ \int_t^s U(c(u)) + \psi(u) + \lambda(u)U(Z(u))du\big| s \in [t, T]\big\}$ *is uniformly integrable under the worst-case scenario;*

3. $\big\{ \int_t^s U(c(u)) - \psi(u) + \lambda(u)U(Z(u))du\big| s \in [t, T]\big\}$ *is uniformly integrable under the best-case scenario;*

4. $\big\{ \int_t^s V_w(u, W(u), \mathcal{X}(u))\pi(u)\sigma(u)W(u)dB^{\mathbb{Q}}(u)\big| s \in [t, T]\big\}$ *is a local martingale under the worst-case and best-case scenarios.*

Then, the optimal strategy $u^*(t) = (\pi^*(t), c^*(t), p^*(t)) \in \mathcal{A}$ is given by (3.14), and the worst-case and the best-case measures are given by (3.16) and (3.15), respectively.

Proof. See Appendix B.3. ■

Theorem 3.3.4 *For the robust optimal control problem (3.10), $u^*(t) = (\pi^*(t), c^*(t), p^*(t)) \in \mathcal{A}$ given by (3.14) is the robust optimal strategy, and $(\underline{\theta}^*(t), \underline{\phi_{ij}^*}(t))$ and $(\overline{\theta}^*(t), \overline{\phi_{ij}^*}(t))$ given by (3.16) and (3.15) are the worst-case measure and the best-case measure, respectively.*

Proof. See Appendix B.4. ■

The verification theorem confirms that the value function of the optimal control problem corresponds to a unique solution of the HJB equation system, leading to the corresponding optimal investment, consumption, and life insurance purchase strategies that can be expressed in (3.14).

3.3.3 Utility Loss

When the wage earner follows a non-robust optimal life insurance purchase and consumption-investment strategy, utility loss will occur. We consider that the wage earner does not adopt the robust optimal strategy $u^*(t) = (\pi^*(t), c^*(t), p^*(t))$ given in Theorem 3.3.1, but adopts a sub-optimal strategy $\widetilde{u}^*(t) = (\widetilde{\pi}^*(t), \widetilde{c}^*(t), \widetilde{p}^*(t))$ instead as if model uncertainty is absent. Suppose that the wage earner ignores model uncertainty. Then the wealth process under the probability measure $\mathbb{P}$ is described by equation (3.2).

Denote the value function by

$$\widetilde{V}(t,w,e_i) = \sup_{(\tilde{\pi},\tilde{c},\tilde{p})\in\mathcal{A}} \mathbb{E}\Bigg\{ \int_t^T e^{-\int_t^s \delta(v)dv}\Big[\bar{F}(s,t)U(c(s)) + f(s,t)\xi(s)U(Z(s))\Big]ds \\ + e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)\zeta(T)U(W(T))\Bigg\}.$$

The corresponding HJB equation is given by

$$\widetilde{V}_t - \lambda(t)\widetilde{V}(t,w,e_i) + \sup_{(\widetilde{\pi},\widetilde{c},\widetilde{p})\in\mathcal{A}} \Bigg\{ \frac{1}{2}\widetilde{\pi}^2\sigma_i^2 w^2 \widetilde{V}_{ww} + (\mu_i - r)\widetilde{\pi} w \widetilde{V}_w + \Big[rw - \widetilde{c} - \widetilde{p} + \iota(t)\Big]\widetilde{V}_w \\ - \delta\widetilde{V} + U(\widetilde{c}) + \lambda(t)\xi_i U(z) + \Big\langle \mathcal{Q}e_i, \widetilde{\mathbf{V}}(t,w)\Big\rangle \Bigg\} = 0, \tag{3.19}$$

where $\widetilde{\mathbf{V}}(t,w) := \big(\widetilde{V}(t,w,e_1), \widetilde{V}(t,w,e_2), \cdots, \widetilde{V}(t,w,e_N)\big)'$, and $\widetilde{V}_t$, $\widetilde{V}_w$, and $\widetilde{V}_{ww}$ denote partial derivatives $\frac{\partial\widetilde{V}}{\partial t}$, $\frac{\partial\widetilde{V}}{\partial w}$, and $\frac{\partial^2\widetilde{V}}{\partial w^2}$, respectively.

Theorem 3.3.5 *If the wage earner ignores the model uncertainty, when $\mathcal{X}(t) = e_i$, the solution of the HJB equation (3.19) is given by*

$$\widetilde{V}(t,w,e_i) = \widetilde{k}(t,e_i)\frac{[w+\widetilde{g}(t)]^{1-\gamma}}{1-\gamma}, \qquad \gamma > 0 \quad \textit{and} \quad \gamma \neq 1,$$

the suboptimal investment-consumption and life insurance strategy is given by

$$\begin{cases} \widetilde{\pi}^*(t) = \frac{\mu_i - r}{\sigma_i^2}\frac{w+\widetilde{g}(t)}{w\gamma}, \\ \widetilde{c}^*(t) = \widetilde{k}^{-\frac{1}{\gamma}}(t,e_i)[w+\widetilde{g}(t)], \\ \widetilde{p}^*(t) = \xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\widetilde{k}^{-\frac{1}{\gamma}}(t,e_i)\eta^{1-\frac{1}{\gamma}}(t)[w+\widetilde{g}(t)] - w\eta(t), \end{cases} \tag{3.20}$$

where $\widetilde{k}(t,e_i)$ and $\widetilde{g}(t)$ satisfy the following equations:

$$\frac{d\widetilde{k}(t,e_i)}{dt} + \Big[1 + \xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\eta^{1-\frac{1}{\gamma}}(t)\Big]\gamma\widetilde{k}^{1-\frac{1}{\gamma}}(t,e_i) + \sum_{j=1}^{N} q_{ij}\widetilde{k}(t,e_j) = a_i(t)\widetilde{k}(t,e_i),$$

where

$$a_i(t) := \delta_i + \lambda(t) + \frac{1}{2}\frac{(\mu_i - r)^2}{\sigma_i^2}\frac{1-\gamma}{\gamma} - r(1-\gamma) - \eta(t)(1-\gamma),$$

and $\widetilde{g}(t) = g(t)$.

Proof. This proof is similar to that of Theorem 3.3.1, and hence we omit it here. ■

The wage earner following a suboptimal strategy will incur a utility loss. Under the suboptimal investment-consumption and life insurance purchase rule, the value function associated with $\widetilde{u}^*(t) = (\widetilde{\pi}^*(t), \widetilde{c}^*(t), \widetilde{p}^*(t))$ is defined by

$$\begin{aligned} \check{V}(t,w,e_i) = \inf_{(\theta,\phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg\{ & \int_t^T \alpha(t)e^{-\int_t^s \delta(v)dv}\Big[\bar{F}(s,t)U(c(s)) + \bar{F}(s,t)\psi(s) + f(s,t)\xi(s)U(Z(s))\Big]ds \\ & + \alpha(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)\zeta(T)U(W(T))\Bigg\} \\ & + \sup_{(\theta,\phi)\in\Theta} \mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg\{ \int_t^T \hat{\alpha}(t)e^{-\int_t^s \delta(v)dv}\Big[\bar{F}(s,t)U(c(s)) - \bar{F}(s,t)\psi(s) + f(s,t)\xi(s)U(Z(s))\Big]ds \\ & + \hat{\alpha}(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)\zeta(T)U(W(T))\Bigg\}, \end{aligned}$$

and the corresponding HJB equation becomes

$$\begin{aligned} & \inf_{(\theta,\phi)\in\Theta}\left\{\alpha_i\left[\mathcal{L}^{(\widetilde{\pi}^*,\widetilde{c}^*,\widetilde{p}^*;\theta,\phi)}\check{V}(t,w,e_i) + U(\widetilde{c}^*) + \psi(t) + \lambda(t)\xi_i U\Big(w + \frac{\widetilde{p}^*}{\eta(t)}\Big)\right]\right\} \\ & + \sup_{(\theta,\phi)\in\Theta}\left\{\hat{\alpha}_i\left[\mathcal{L}^{(\widetilde{\pi}^*,\widetilde{c}^*,\widetilde{p}^*;\theta,\phi)}\check{V}(t,w,e_i) + U(\widetilde{c}^*) - \psi(t) + \lambda(t)\xi_i U\Big(w + \frac{\widetilde{p}^*}{\eta(t)}\Big)\right]\right\} = 0, \end{aligned} \tag{3.21}$$

where

$$\begin{aligned}\mathcal{L}^{(\widetilde{\pi}^*,\widetilde{c}^*,\widetilde{p}^*;\theta,\phi)}\check{V}(t,w,e_i) :=& \check{V}_t + \Big[(\mu_i - r)\widetilde{\pi}^* w + rw - \widetilde{c}^* - \widetilde{p}^* + \iota(t) + \theta\widetilde{\pi}^*\sigma_i w\Big]\check{V}_w \\ &+ \frac{1}{2}\widetilde{\pi}^{*^2}\sigma_i^2 w^2 \check{V}_{ww} - \delta_i \check{V} - \lambda(t)\check{V} + \langle \mathcal{Q}^{\phi} e_i, \check{\mathbf{V}}(t,w)\rangle,\end{aligned} \tag{3.22}$$

where $\check{\mathbf{V}}(t,w) := (\check{V}(t,w,e_1), \check{V}(t,w,e_2), \ldots, \check{V}(t,w,e_N))$.

The value function $\check{V}$ is given by

$$\check{V}(t,w,e_i) = k(t,e_i)\frac{[w+g(t)]^{1-\gamma}}{1-\gamma}, \qquad \gamma > 0 \quad \text{and} \quad \gamma \neq 1.$$

Denote by $(\underline{\theta}^{\diamond}, \underline{\phi_{ij}^{\diamond}})$ and $(\overline{\theta}^{\diamond}, \overline{\phi_{ij}^{\diamond}})$ the worst-case and best-case distortion processes, respectively. By the first-order condition with respect to θ and ϕ_{ij}, we get

$$\begin{cases} \underline{\theta}^{\diamond}(t) = -\frac{u_i - r}{\sigma_i}\frac{\beta_i}{\gamma}, \\ \underline{\phi_{ij}^{\diamond}}(t) = \exp\left\{\frac{\beta_i}{1-\gamma}\left[1 - \frac{k(t,e_j)}{k(t,e_i)}\right]\right\}, \end{cases} \tag{3.23}$$

and

$$\begin{cases} \overline{\theta}^{\diamond}(t) = \frac{u_i - r}{\sigma_i}\frac{\beta_i}{\gamma}, \\ \overline{\phi_{ij}^{\diamond}}(t) = \exp\left\{-\frac{\beta_i}{1-\gamma}\left[1 - \frac{k(t,e_j)}{k(t,e_i)}\right]\right\}. \end{cases} \tag{3.24}$$

Substituting (3.20), (3.23), and (3.24) into the HJB equation (3.21), we get

$$\begin{aligned}&\frac{dk(t,e_i)}{dt} + \Big[1 + \xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\eta^{1-\frac{1}{\gamma}}(t)\Big]\gamma k^{1-\frac{1}{\gamma}}(t,e_i) + \frac{1-\gamma}{\beta_i}\Bigg\{\alpha_i \sum_{j=1,j\neq i}^{N} q_{ij}\Big(\underline{\phi_{ij}^{\diamond}}(t)\log\underline{\phi_{ij}^{\diamond}}(t) - \underline{\phi_{ij}^{\diamond}}(t) + 1\Big) \\ &- \hat{\alpha}_i \sum_{j=1,j\neq i}^{N} q_{ij}\Big(\overline{\phi_{ij}^{\diamond}}(t)\log\overline{\phi_{ij}^{\diamond}}(t) - \overline{\phi_{ij}^{\diamond}}(t) + 1\Big)\Bigg\}k(t,e_i) + \sum_{j=1}^{N} q_{ij}k(t,e_j)\Big[\alpha_i\underline{\phi_{ij}^{\diamond}}(t) + \hat{\alpha}_i\overline{\phi_{ij}^{\diamond}}(t)\Big] \\ &- c_i(t)k(t,e_i) = 0,\end{aligned}$$

where

$$c_i(t) := \delta_i + \lambda(t) + \frac{1}{2}\frac{(\mu_i - r)^2}{\sigma_i^2}\frac{1-\gamma}{\gamma}\Big[1 - \frac{\beta_i(\hat{\alpha}_i - \alpha_i)}{\gamma}\Big] - r(1-\gamma) - \eta(t)(1-\gamma),$$

and $g(t)$ is given by (3.18).

Motivated by P. Wang and Li (2018), D. Li, Zeng, and Yang (2018), and Branger, Larsen, and Munk (2013), we define the utility loss for ignoring model uncertainty as follows:

$$UL(t,e_i) = 1 - \frac{V(t,w,e_i)}{\check{V}(t,w,e_i)} = 1 - \frac{f(t,e_i)}{k(t,e_i)}.$$

As we explained earlier in this chapter, the individual is suspicious about the reference model, therefore she aims to find an alternative model that decreases the risk associated with model misspecification. However, if the individual ignores model uncertainty, she believes fully in the reference model, and therefore takes on a rational strategy. The utility loss reflects the importance of considering model ambiguity. If the individual ignores the model ambiguity, she will suffer a utility loss.

3.4 Numerical Analysis

In this section, we analyze the impacts of model parameters on the robust optimal life insurance purchase and consumption and also investigate the effect of the model uncertainty on utility loss. Throughout we consider a wage earner who starts working at age 25 and retires 35 years later, and whose initial wage at age 25 is $\$40,000$, growing at a rate of 3% per year.

In our numeric examples, we consider two states Markov chain $\mathcal{X}$, that is, State 1 and State 2, representing a bear market and bull market, respectively. The rate matrix of the Markov chain is given by

$$\begin{pmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}.$$

Suppose that the configurations of the parameter values are listed in Table 3.1. Since State 1 and State 2 represent the bear market and bull market, respectively, the expected return $\mu(t, e_2)$ in the bull market is higher than the expected return $\mu(t, e_1)$ in the bear market. Furthermore, the risky shares are more volatile in State 1 than in State 2.

The hazard rate $\lambda(t)$ and the premium-insurance ratio $\eta(t)$ are defined as follows:

1. Hazard rate (Gompertz hazard function) $\lambda(t)$: $\lambda(t) = 0.001 + e^{-9.5+0.1\times t}$;

2. Premium-insurance ratio $\eta(t)$: $\eta(t) = 0.001 + e^{-9.5+0.1\times t}$.

Remark 3.4.1 *Note that the parameters $\lambda(t)$ and $\eta(t)$ have a very small value in the real-world data. The life insurance company set up the premium-insurance ratio $\eta(t)$, which is greater than the hazard rate $\lambda(t)$ in order to make a profit. That is because the expected profit*

Table 3.1: Value of Parameters

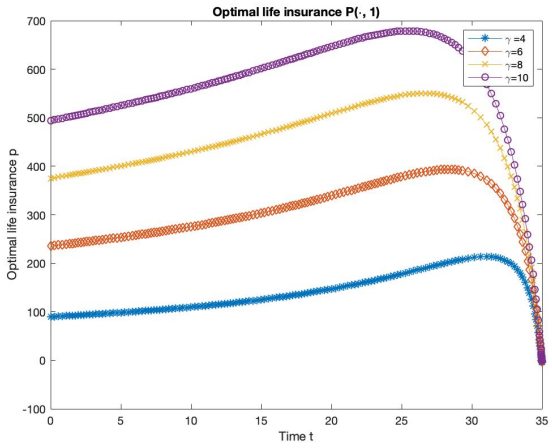
	$\mu(t,\cdot)$	$\sigma(t,\cdot)$	$\delta(t,\cdot)$	$\alpha(t,\cdot)$	$\beta(t,\cdot)$	$\xi(t,\cdot)$	$\zeta(t,\cdot)$
State 1	0.09	0.19	0.03	0.9	14	1	1.5
State 2	0.15	0.13	0.03	0.6	14	1.7	3
	W	r	γ	T			
	10000	0.03	4	35			

rate of the life insurance is $p(t)\frac{\eta(t)-\lambda(t)}{\eta(t)}$. *Thus, life insurance makes a profit when* $\eta(t) > \lambda(t)$. *In fact, life insurance is fair due to the expected profit rate being zero. Therefore, we set up* $\eta(t) = \lambda(t)$ *for our numerical experiments.*

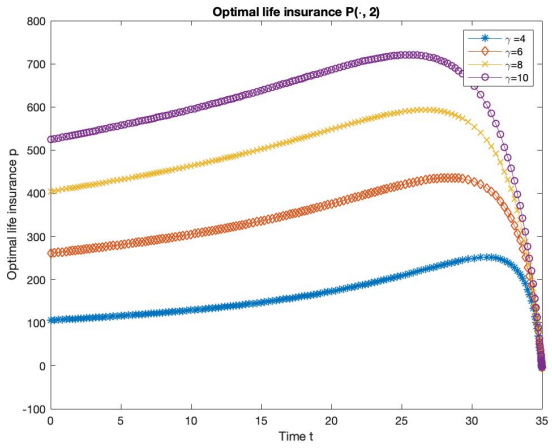
3.4.1 Effects of Model Parameters on Robust Optimal Strategies

In this subsection, we focus on the impacts of model parameters including risk aversion, interest rate, discount rate, mortality rate, and ambiguity aversion on the robust optimal life insurance purchase and consumption strategies.

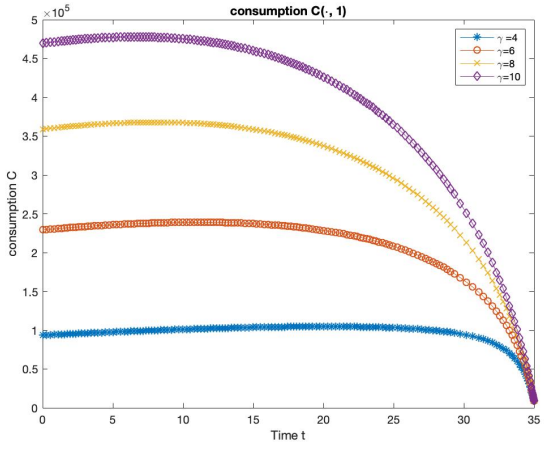
Figure 3.1 shows the sensitivity of the robust optimal life insurance purchase and consumption on the risk aversion parameter γ. It can be seen from Figure 3.1 that γ has a positive effect on the wage earner's life insurance purchase and consumption. As γ increases, the wage earner becomes more risk-averse. The more risk-averse individual is prone to invest less in the stock market, and reallocate their wealth to life insurance and consumption when they are young. The reasonable explanation is that the wage earner has a lower mortality risk when they are young; compared to the welfare of risk-free assets and terminal wealth, life insurance and consumption create more of a sense of well-being for the wage earner. In addition, the wage-earner has a higher mortality rate when they are old. The wage earner is willing to purchase more life insurance to cover the mortality risk and consume less to save wealth to gain a sense of well-being from terminal wealth because the terminal wealth has more weight than the total legacy. Thus, risk-aversion has little impact on life insurance and consumption when the wage earner ages.



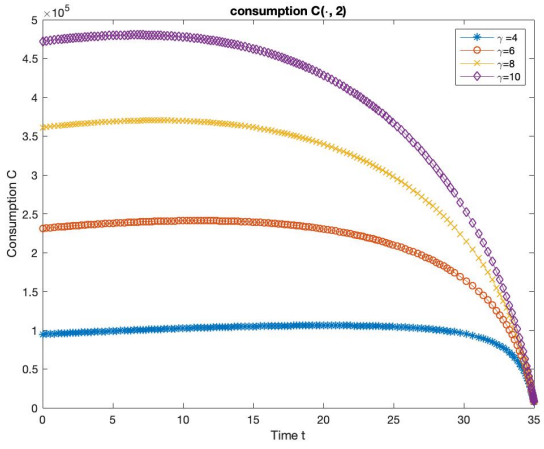
(a)



(b)

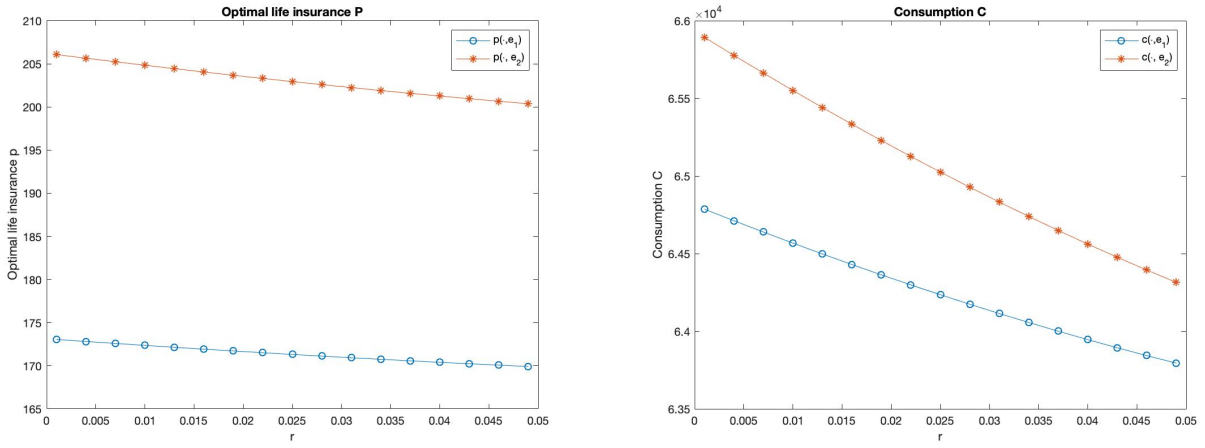


(c)



(d)

Figure 3.1: Sensitivity plots on robust optimal life insurance purchase and consumption in response to change in risk aversion γ.

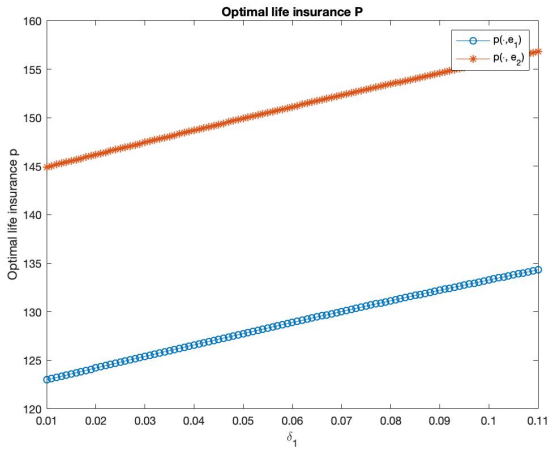


(a) (b)

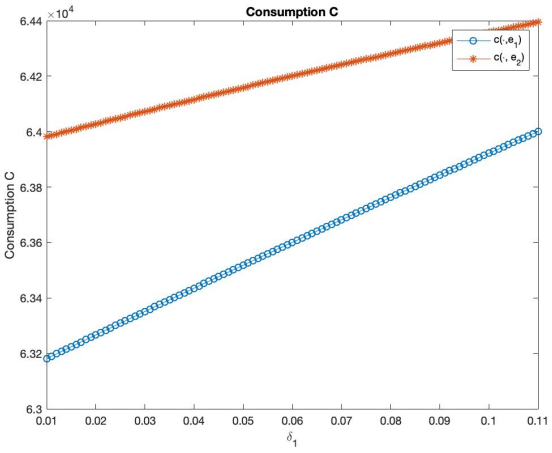
Figure 3.2: Sensitivity plots on robust optimal life insurance purchase in response to change in interest rate r.

Figure 3.2 reveals the sensitivity of the robust optimal life insurance purchase and consumption on the interest rate r. From this figure, we obtain that life insurance and consumption decrease with respect to interest rate r. With the increases of r, the risk-free asset is more attractive to the wage earner, who is inclined to shift more wealth to the risk-free asset. As a result, the wage earner reduces consumption and purchases less life insurance.

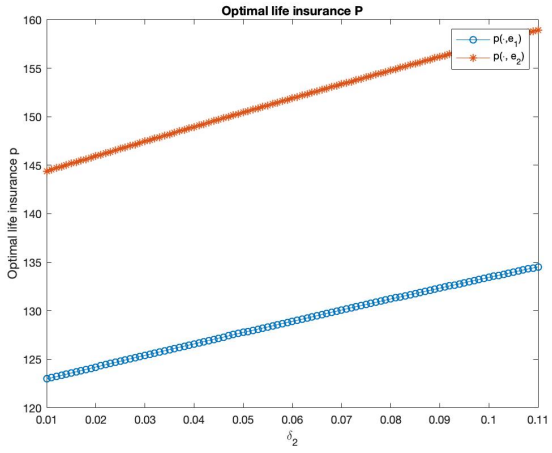
Figure 3.3 discloses the sensitivity of the optimal life insurance and consumption on discount rate δ_1 and δ_2, respectively. As shown in subfigures 3.3(a) and 3.3(b), the utility discount rate δ_1 has a positive effect on the life insurance purchase and consumption for the wage earner in both the bull market and the bear market, respectively. As δ_1 increases, the wage earner has a higher time preference. The higher the time preference, the higher the discount placed on costs payable in the future. That is, the individual with a high time preference is focused substantially on their well-being in the present and the immediate future; in contrast, the individual with a low time preference places more emphasis on their well-being in the distant future. Thus, the wage earner with a higher discount factor tends to allocate more wealth to life insurance and consumption to improve their well-being in the present. For the same reason, subfigures 3.3(c) and 3.3(d) disclose that δ_2 also has positive effects on life insurance and consumption in both the bull market and the bear



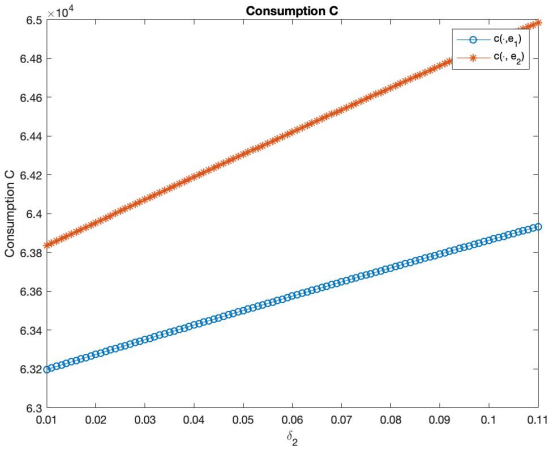
(a)



(b)



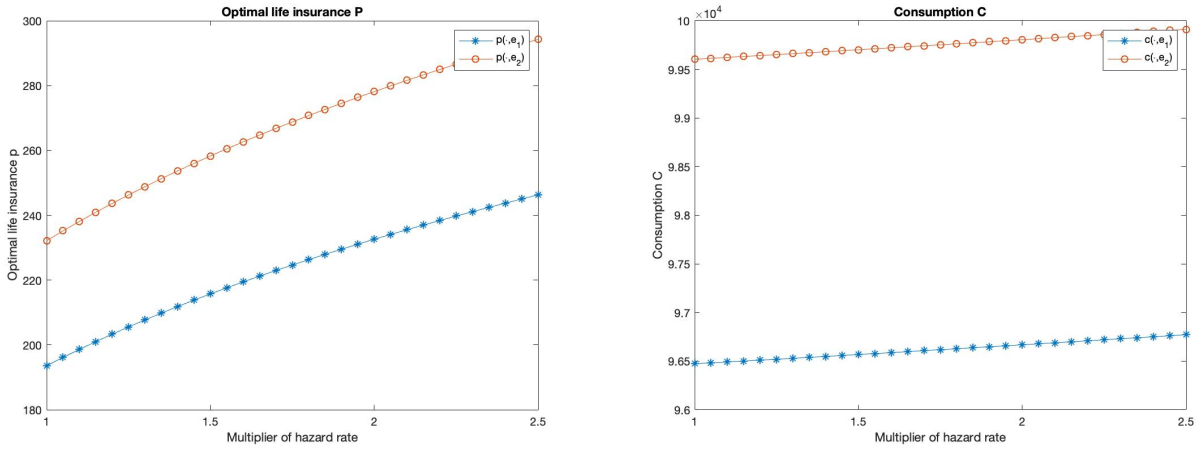
(c)



(d)

Figure 3.3: Sensitivity plots on robust optimal life insurance purchase and consumption in response to change in utility discount factor δ_1 and δ_2.

market. In addition, subfigures 3.3(a) and 3.3(b) highlight that the optimal life insurance $p^*(\cdot, e_1)$ and consumption $c^*(\cdot, e_1)$ in the bear market are more sensitive to the change in discounted rate δ_1 in the bear market. Subfigures 3.3(c) and 3.3(d) present that optimal life insurance $p^*(\cdot, e_2)$ and consumption $c^*(\cdot, e_2)$ in the bull market are more sensitive to the change in discounted rate in the bull market δ_2. One possible explanation for this is that the regime states (macroeconomic conditions) dominate the main effect of the discounted rate on optimal life insurance and consumption. That is, in a bear market, δ_1 has more of an effect on optimal life insurance $p^*(\cdot, e_1)$ and consumption $c^*(\cdot, e_1)$ compared to δ_2; in contrast, under a bull market, δ_2 has more of an effect on optimal life insurance $p^*(\cdot, e_2)$ and consumption $c^*(\cdot, e_2)$ compared to δ_1.



(a) (b)

Figure 3.4: Sensitivity plots on robust optimal life insurance purchase and consumption in response to change in hazard rate λ.

Figure 3.4 displays the sensitivity of the robust optimal life insurance purchase and consumption on hazard rate $\lambda(t)$. To simplify the explanation, we first introduce the multiplier R of $\lambda(t)$, the hazard rate $\lambda^{'}(t) = \lambda(t) \times R$. When $R = 1$, the hazard rate is at the original setting; when $R > 1$, the hazard rate is higher than the original. Therefore, with the higher multiplier R, the wage earner has a higher mortality risk. To protect their family, the wage earner is inclined to purchase more life insurance to cover the higher mortality risk. In addition, from Figure 3.4(b), we obtain that the mortality rate has positive effects on the optimal

consumption. One possible explanation is that the wage earner is inclined to reduce their savings for the future and consume more to gain a sense of well-being in the present due to bad health conditions.

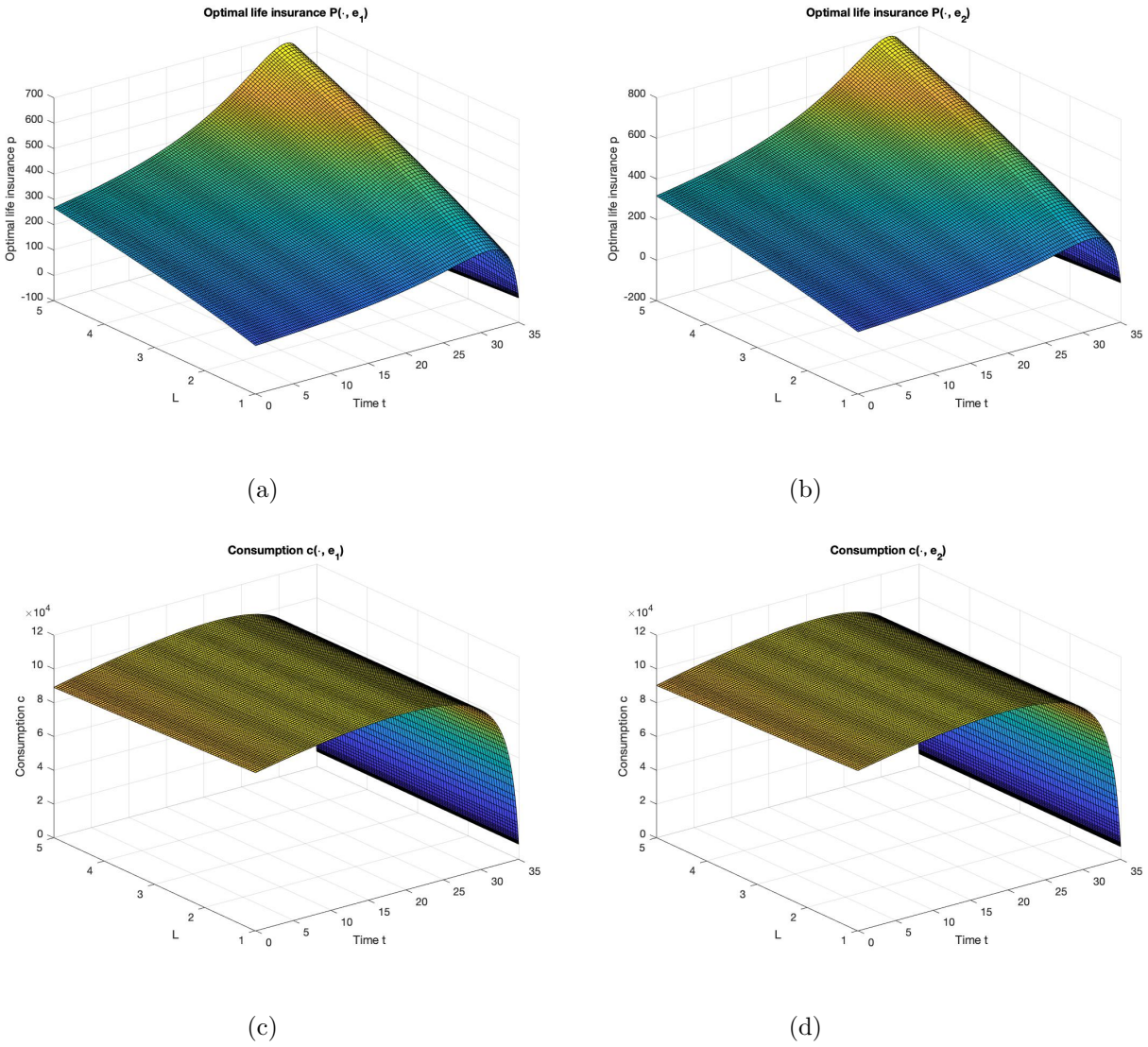


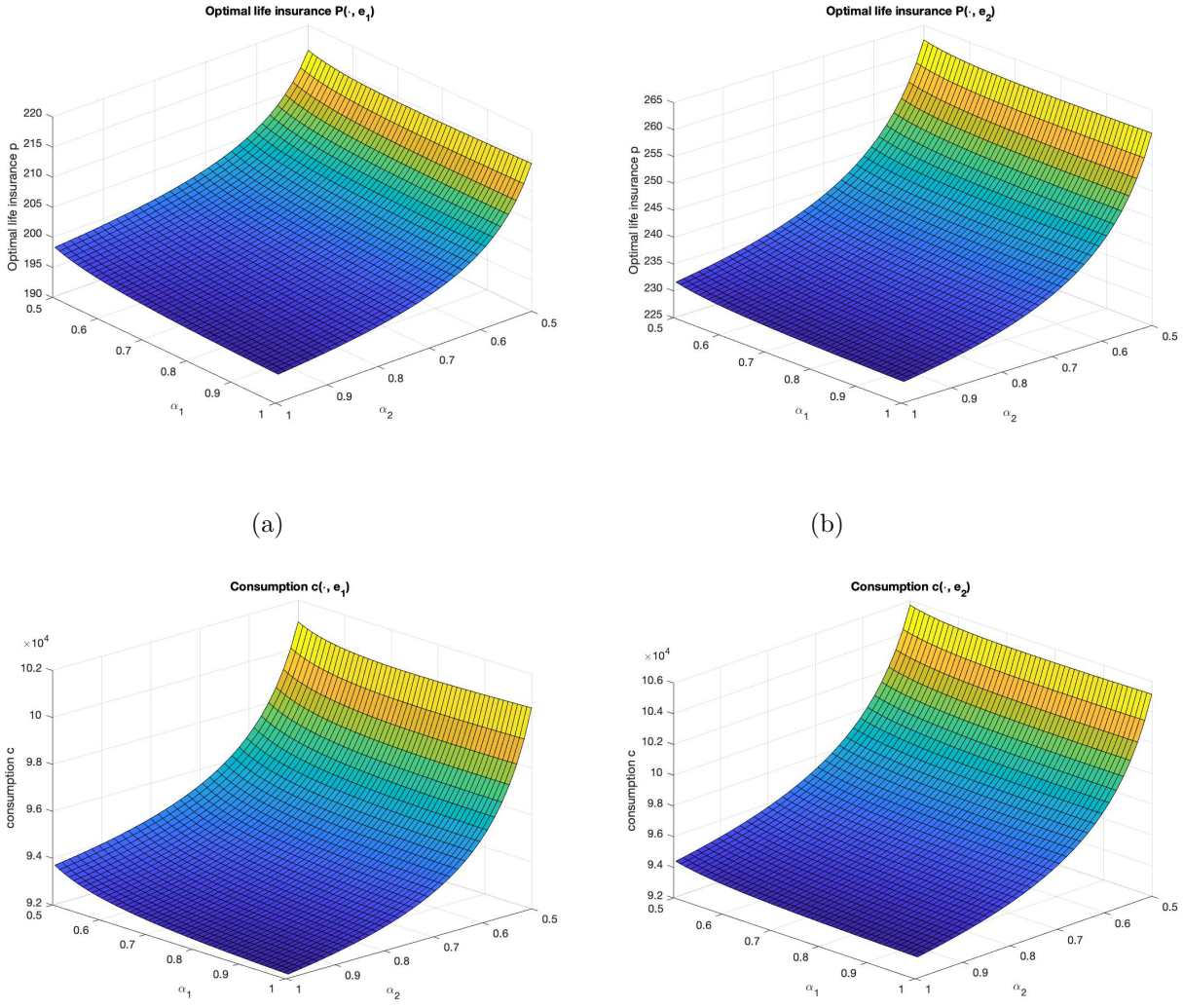
Figure 3.5: Sensitivity plots on robust optimal life insurance purchase and consumption in response to change in Loading factor L.

In Figure 3.5, we vary the loading factor of $\eta(t)$, which is defined as $L = \frac{\eta(t)}{\lambda(t)}$, from 1 to 5. The larger the loading factor, the smaller the legacy the wage earner will receive when they die. To protect their family, the wage earner will have to pay a greater premium for the same amount of legacy. That is, the higher loading factor makes the life insurance policy

more expensive for the wage earner. When the wage earner is young, they are willing to pay more premium at the beginning if the insurance premium is slightly more expensive; however, if it is too expensive to maintain the insurance policy, they may give up the life insurance protection. Moreover, when the wage earner is near retirement age, they will have a higher mortality rate and will be willing to pay a higher premium even when the insurance is more expensive. In addition, Figure 3.5 illustrates the effects of loading factor on the optimal consumption. As life insurance is more expensive, the wage earner has to reduce the consumption to make up the expenditure of life insurance coverage.

Figure 3.6 shows the impacts of the ambiguity-aversion parameters α_1 and α_2 on robust optimal life insurance and consumption in the bull market and bear market, respectively. We can see that the ambiguity aversion's parameters α_1 and α_2 have negative effects on optimal life insurance and consumption for both markets. The analysis on robust optimal life insurance behavior is rather complicated since it depends on other macroeconomic factors. In subfigures 3.6(a) and 3.6(c), with higher levels of ambiguity aversion, the wage earner cares more about the worst-case scenario. That is, the wage earner is more pessimistic about the future and is inclined to seek a more conservative strategy. Thus, they are also inclined to save more for the future in the bear market by reducing life insurance purchase and consumption to alleviate stock market and insurance risk. As shown in 3.6(b) and 3.6(d), optimal life insurance and consumption in the bull market decrease with respect to α_1 and α_2 for the same reason. Surprisingly, we notice that optimal life insurance and consumption under the bear and bull markets are more sensitive to the change in α_2 rather than α_1. One possible explanation for this finding is that the rational wage earner pays more attention to the model uncertainty in the bear market. That is, the rational wage earner suspects more about the reference model and aims to seek a more robust model in the bear market. Thus, we can conclude that optimal life insurance purchase and consumption are more sensitive to the change in δ_2 compared to the change of δ_1.

Overall, our numeric results illustrate the sensitivity of optimal life insurance and consumption to various parameters. In addition, as shown in Figure 3.1-3.5, optimal life insurance and consumption are higher (resp. lower) when the wage earner expects regime-switching in a bull market (resp. a bear market). Since the wage earner is optimistic about



(a) (b)

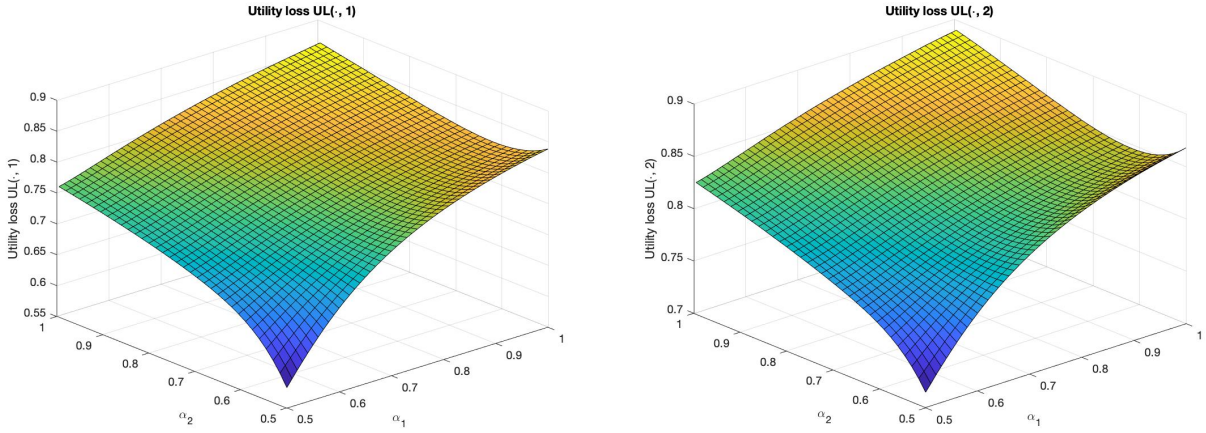
(c) (d)

Figure 3.6: Sensitivity plots on robust optimal life insurance purchase and consumption in response to change in ambiguity aversion α_1 and α_2.

the future under the bull market, they are more inclined to undertake risk when the regime-switching market is in the bull market. Thus, they are tempted to allocate more wealth to life insurance and consumption in the bull market. Moreover, our results also highlight that the effect of regime-switching on robust optimal life insurance and consumption is strong when the wage earner is young, while the effect is negligible when the wage earner is near the retirement time T. The reasonable explanation is that the probability of regime-switching from one state to another state decreases when the wage earner is nearly at retirement age.

3.4.2 Effects of Model Parameters on Utility Loss

In this subsection, we illustrate the utility loss for the wage earner who ignores the model uncertainty by numerical experiment and analyzes the impacts of model parameters on utility loss.

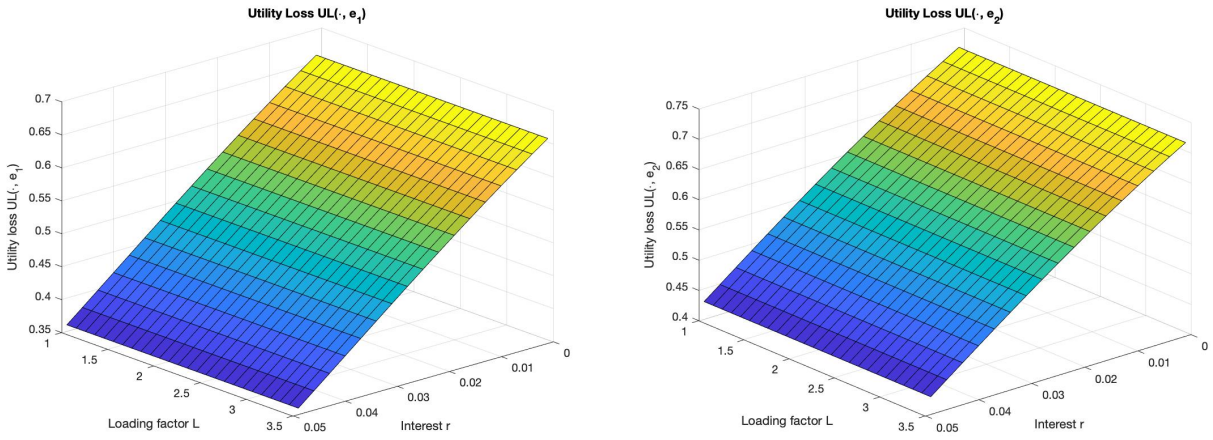


(a) (b)

Figure 3.7: Effects of ambiguity aversion α_1 and α_2 on the utility loss.

Figure 3.7 depicts the effects of ambiguity aversion α_1 and α_2 on utility loss. We find that the utility loss increases with respect to ambiguity aversion α_1 and α_2, respectively. The more ambiguity-averse the wage earner is, the more suspicion about the reference model the wage earner has. Therefore, they are motivated to seek a more conservative portfolio strategy to alleviate uncertainty. Indeed, when the wage earner is more uncertain about

the reference model, ignoring the model uncertainty may suffer a greater utility loss. In addition, comparing the two subfigures 3.7(a) and 3.7(b), the utility loss in the bear market $UL(\cdot, e_1)$ is more sensitive to α_1, while that of bull market $UL(\cdot, e_2)$ is more sensitive to α_2. When the wage earner is more concerned about the worst case in the bear market, they suspect more about the reference model, which leads to more model uncertainty. Therefore, if uncertainty is ignored, ambiguity aversion in the bear market α_1 will lead to a higher utility loss. For the same reason, utility loss in the bull market is more sensitive to α_2 than α_1. In summary, it is necessary to consider ambiguity aversion in the optimal life insurance and investment-consumption problem.



(a) (b)

Figure 3.8: Effects of interest r and loading factor L on utility loss.

Figure 3.8 illustrates the effects of the loading factor L and the interest rate r on the utility loss. It shows that the utility loss decreases with respect to the loading factor L and the interest rate r. As r increases, the risk-free asset is more attractive to the wage earner. Thus, the wage earner is inclined to allocate more wealth in the risk-free asset, and reduce investment in the stock market, which leads to a reduction of the uncertainty in the stock market. If the uncertainty of the model is ignored, it will lead to a smaller utility loss. For the same reason, with the higher loading factor L, life insurance becomes more expensive for the wage earner. The wage earner allocates more wealth to cover expenditure for life insurance. Hence, they have less uncertainty due to lower stock investment. That is, there

is lead to less utility loss if the wage earner ignores the model uncertainty.

Generally speaking, these figures show that if the uncertainty of the model is ignored, the utility loss will occur. We obtain that ambiguity aversion plays a pivotal role in utility loss. If we assume that the wage earner only holds the extremely ambiguity-averse attitude in the worst-case scenario, accepting the suboptimal strategy may lead to higher utility loss.

3.5 Conclusion

In this chapter, we study the robust optimal life insurance and investment-consumption problem for the wage earner who is endowed with initial wealth and receives income over a random lifetime. Taking into account the worst-case scenario and the best-case scenario, our proposed regime-switching alpha-ambiguity expected discounted utility form allows the wage earner to have different levels of ambiguity aversion in different economic states. By solving the HJB equations, we obtain a robust strategy to the problem under regime-switching alpha-ambiguity utility. From the numerical experiment and theoretical analyses, it shows that ambiguity aversion has negative effects on optimal life insurance and consumption in both a bear market and a bull market. Compared with the bull market, the rational wage earner is more concerned about uncertainty in the bear market. In addition, considering ambiguity aversion will lead to less utility loss.

We explore the derivation of a pre-committed strategy for the robust optimal life insurance and investment-consumption problem. Such a strategy underscores a proactive and resolute commitment to a pre-defined course of action, undeterred by any subsequent shifts in the environment. While we've focused on the pre-committed approach in this work, we acknowledge the importance of time-consistent strategies and reserve an in-depth exploration of this topic for future research.

4 Robust and Risk-sensitive Markov Decision Process with Hidden Regimes

In this chapter, we explore the complexities of decision-making with hidden regimes, focusing on the financial markets. By incorporating hidden regimes into Markov Decision Processes (MDPs) and leveraging Reinforcement Learning (RL), we aim to offer a robust framework that enhances portfolio performance and risk management in unpredictable market conditions.

4.1 Introduction

Sequential decision-making is a fundamental problem in many real-world applications where decision-makers are tasked with making a sequence of decisions over time to achieve a specific goal. Such problems are widespread across various fields, including finance, healthcare, robotics, and control engineering. For instance, in finance, portfolio managers need to make a sequence of investment decisions to maximize returns while minimizing risk. Markov Decision Process (MDP) is a widely used framework to model the sequential decision-making problems. However, due to the complex and dynamic nature of financial markets, the transition probabilities between states in the MDP framework are often unknown and uncertain, making it challenging to develop accurate models for informed decision-making. This necessitates the development of effective and robust sequential decision-making models for portfolio management problems.

To tackle this issue, researchers have developed various approaches to improve the robustness of sequential decision-making models. One widely adopted approach is robust dynamic programming (DP), which is a mathematical framework for solving sequential decision-

making problems by recursively splitting the problem into smaller subproblems. This approach allows the decision-maker to efficiently compute a beneficial overall strategy based on the information state. However, the computational effort required to compute the optimal policy for a robust DP approach can be overwhelming in portfolio management problems due to the large number of states involved. This obstacle has been addressed by Bertsekas and Tsitsiklis (1996); De Farias and Van Roy (2003), who provide the efficient approximation algorithms to compute the optimal solutions. Iyengar (2005); Nilim and El Ghaoui (2005) formulate robust control problems for robust MDPs that model an uncertainty set of transition probabilities and derive an optimal policy that performs well under worst-case scenarios using robust dynamic programming. Their contribution informs subsequent research by Wiesemann et al. (2013), who propose a robust Markov decision process formulation to address the issue of uncertainty in MDPs, where the transition probabilities are unknown or uncertain. They construct a confidence region for the unknown parameters with a specified probability and determine a policy that maximizes the worst-case performance over this region. In a further development, Goyal and Grand-Clement (2023) present a new approach for solving robust MDPs that go beyond the traditional rectangular models. They propose a novel algorithm for solving these extended robust MDP models, called "A factor matrix uncertain model". This algorithm considers a factor model for probability transitions, where the transition probability is a linear function of a factor matrix that is uncertain and belongs to a factor matrix uncertainty set. The algorithm provides a fairly general model of uncertainty in probability transitions, allowing the decision-maker to capture dependence between probability transitions across different states, and it is significantly less conservative than prior approaches.

A major drawback of previous work on robust MDPs is that they all focused on the planning problem with no effort to learn the uncertainty. Since it is often difficult to accurately quantify the uncertainty in practice, the solutions to the robust MDP can be conservative if a too large uncertainty set is utilized. Recent advancements in deep Reinforcement Learning (RL) have expanded the robust MDP framework to tackle large-scale problems with high-dimensional state and action spaces. This development has achieved great success in resolving diverse decision-making problems in finance. However, real-world applications often feature

environments characterized by uncertainty and risk, which can lead to the poor performance or failure of RL algorithms. To address this challenge, RL with risk-sensitive objective has emerged as a critical research area, aiming to develop algorithms that can effectively handle uncertainties and risks in complex environments. Building on this line of research, Tamar, Di Castro, and Mannor (2012) propose a framework for local policy gradient style algorithms in RL for variance-related risk criteria. Their novel RL algorithm involve risk criteria by optimizing both the expected cost and the variance of the cost. The practical applicability of this approach in a portfolio management problem further underscores its significance. Furthering this exploration of risk-sensitive objective, Tamar, Glassner, and Mannor (2015) implement RL with risk measure CVaR to capture the expected loss beyond a certain threshold, and propose a novel approach for computing the gradient of CVaR in the form of a conditional expectation using a sampling-based optimization method. Chow et al. (2015) employ a risk-sensitive CVaR as objective function to replace a standard risk-neutral expectation, and show that a CVaR objective can capture the risk sensitivity. Lastly, Tamar et al. (2016) propose a novel risk-sensitive objective function for RL that considers the consequences of different decisions in a coherent manner. They propose a sampling-based algorithm for estimating the gradient of coherent risk. This approach is further validated through various simulation scenarios, including portfolio management and control engineering applications, thereby offering a promising direction for enhancing the efficacy of RL in uncertain environments.

Financial markets are characterized by hidden regimes that must be inferred from observable variables. Failing to consider these regimes can result in suboptimal investment decisions, increased risk exposure, and financial losses. Robust and Risk-sensitive MDPs that incorporate hidden regimes have become powerful tools for addressing these challenges (Levy, Vazquez-Abad, & Costa, 2006; A. Zhang, Sodhani, Khetarpal, & Pineau, 2020; Y. Zhang & Desilva, 2014). These approaches involve finding a policy that maximizes the worst-case performance over all possible transition probabilities, which is governed by hidden regimes. Through the incorporation of transition probability uncertainties and the formulation of the problem as risk-sensitive MDPs, decision-makers can create policies that are robust to market condition changes, thereby enhancing portfolio performance and risk management. In this study, we initially develop a dynamic programming (DP) framework for both finite and

infinite horizon robust MDPs with hidden regime rules. Inspired by the work of Tamar et al. (2016), we adopt a risk-sensitive reinforcement learning (RL) objective by using a risk envelope as the uncertainty set of transition probabilities, which is associated with hidden regimes. This approach allows decision-makers to take into account uncertainty in the transition probabilities and constructive policies that are robust to changes in the market conditions. Utilizing the risk envelope as the uncertainty set enables the RL algorithm to hedge effectively against uncertainty and the risk of making suboptimal decisions, leading to improved performance in practice.

This study provides three key contributions in the realm of sequential decision-making under uncertainty in systems with hidden regime rules: first, we introduce an innovative framework that seamlessly integrates hidden regime rules into both finite and infinite horizon robust Dynamic Programming (DP). This integration permits the incorporation of uncertainties associated with these regimes into the decision-making process, enabling the formulation of robust strategies. By selecting a particular uncertainty, we can obtain a statistically precise representation of uncertainty and solve the robust problem through classical recursion. This unique approach demonstrates practical efficiency and applicability in addressing challenges related to uncertainty in complex systems. Second, we demonstrate that when the set of conditional measures of transition probabilities associated with hidden regime rules, satisfies the rectangularity property, most of the key findings in DP theory, such as the Bellman recursion, the optimality of deterministic Markov policies, and the contraction property of the value iteration operator, can be applied to natural robust scenarios. In addition, from RL perspective, we adopt a risk-sensitive objective and construct a risk envelope over transition probabilities to represent the worst-case scenario. To address the risk-sensitive objective with uncertainty, we propose a novel approach to the robust and risk-sensitive MDP by treating the risk envelope of transition probabilities as an uncertainty set and maximizing the worst-case performance over the expectation. By integrating hidden regimes, we improve our capacity to identify potential market risks related to our decisions, and develop robust strategies that adapt to changing market conditions. The effectiveness of our approach is confirmed through empirical results.

The chapter is organized as follows. Section 4.2 briefly reviews the related works of robust

MDP and risk-sensitive MDP. Section 4.3 formulates the robust MDPs with hidden regime rules. Section 4.4 introduces the risk-sensitive RL with risks envelope. Section 4.5 implement our proposed algorithm. Section 4.6 shows the experiment results for real-world data and simulated data. The final conclusion is presented in section 4.7.

4.2 Relative Work

The MDP, pioneered by Puterman (1994), is a mathematical framework used to model and solve sequential decision-making problems in dynamic environments. In the MDP framework, decision-making usually entails minimizing a performance objective that is risk-neutral in nature. However, this approach does not consider the cost's variability (fluctuations around the average) or the impact of modeling errors, which can substantially influence the overall performance. Risk-sensitive MDPs provide a solution to the issue of cost variability by replacing the risk-neutral expectation with a risk-measure of the total discounted cost, that is, variance, Value-at-Risk (VaR), or Conditional-VaR (CVaR). Meanwhile, robust MDPs (Iyengar, 2005; Nilim & El Ghaoui, 2005) tackle the issue of sensitivity to modeling errors by optimizing decisions based on a set of plausible MDP parameters. Robust MDPs provide decision-makers with a policy that is robust and performs reasonably well across a range of possible parameter values under the worst-case scenario.

4.2.1 Risk-sensitive MDPs

Research on risk-sensitive MDPs have been conducted for many years, with initial studies concentrating on exponential utility and mean-variance. Comparing robust MDPs, risk-sensitive MDPs operate under the assumption that parameters are known exactly. The objective of a risk-sensitive MDP is to minimize the value of a risk measure, such as Value-at-Risk, Conditional Value-at-Risk (CVaR). Stella et al. (1998) introduce the risk-sensitive MDP model, where the objective is to find a policy that maximizes the probability that the cumulative cost is within some user-defined cost threshold. They propose a Value Iteration (VI) algorithm to solve the problem. However, their algorithm faces scalability issues, limiting its applicability to large-scale problems. Based on previous work, Hou et al. (2014) revisit

the risk-sensitive MDP model and propose a novel approach to solving the problem, called Topological Value Iteration. The new algorithm is more efficient and faster than the original VI algorithm, addressing some scalability concerns. In a different context, Rockafellar et al. (2000) explore the risk-sensitive objective CVaR for portfolio management problems. They propose an algorithm for minimizing the CVaR, which they demonstrate to be a more robust measure of risk that accounts for extreme losses in the tail of the loss distribution. This approach provides a framework for using CVaR as an alternative risk measure that better captures tail-risk events. This is particularly beneficial in scenarios where decisions need to be made under uncertainty and the potential cost of extremely negative outcomes needs to be minimized. Further extending the field, Borkar and Jain (2014) develop a novel framework for modeling risk-constrained MDPs that measure risk through CVaR. The objective is to find a policy that optimizes a performance criterion while adhering to the given risk constraints. This approach has been proven to asymptotically converge to an optimal risk-constrained policy. The authors also propose a numerical method for solving the resulting optimization problem, which involves a combination of value iteration and linear programming. Building on these advancements, Chow et al. (2015) consider risk-sensitive MDPs with a CVaR objective, referred to as CVaR MDPs. They provide a new optimization algorithm for CVaR MDP, which minimize a risk-sensitive CVaR of the total cost in the CVaR MDP leverages the state augmentation procedure and propose an approximate algorithm with convergence analysis. Furthermore, Chow, Ghavamzadeh, Janson, and Pavone (2017) present efficient reinforcement learning algorithms for risk-constrained MDPs that integrates percentile risk criteria into the standard MDP framework, where risk is represented via a chance constraint or a constraint on the CVaR of the cumulative cost. They provide a policy gradient and actor-critic algorithm for solving the problem, and demonstrate the effectiveness through real-world applications. Recently, Du, Wang, and Huang (2022) propose a new algorithm, called Ïterated CVaR and the Worst Path ẗhat takes into account the trade-off between the expected reward and the risk of taking suboptimal actions. This algorithm employs CVaR as a risk measure and iteratively updates the policy by maximizing a lower bound of the expected CVaR along the worst path. The authors provide theoretical guarantees for the convergence of the algorithm and demonstrate its effectiveness through experiments on a

variety of risk-sensitive MDPs. Furthering this line of research, Rigter, Duckworth, Lacerda, and Hawes (2022) show that the existence of multiple policies that can achieve the optimal CVaR, and this realization motivates for authors to propose a lexicographic approach that minimizes the expected cost while ensuring the CVaR of the total cost remains optimal.

4.2.2 Robust MDPs

The concept of robustness has been extensively studied in optimization and optimal control (Ben-Tal, El Ghaoui, & Nemirovski, 2009; Hansen & Sargent, 2008). Robust MDPs are designed to handle parameter uncertainties, which are situation some parameters cannot be estimated accurately. The objective of robust MDP seeks to find a policy that maximizes the minimum expected total reward for all possible parameter values, considering the fact that the parameter values can fluctuate within an uncertainty set, ensuring robustness against uncertainty and variations in the system. The solution to the robust MDP problem provides a performance guarantee for all uncertain MDP models, thereby offering robustness to model mismatch. Building on this foundation, Lim, Xu, and Mannor (2013) propose a novel algorithm that incorporates robust optimization techniques into the RL framework, allowing for more effective decision-making in situations where there is uncertainty in the parameters of the transition probabilities of the MDP. This literature also provides a theoretical analysis of the proposed algorithm and demonstrates its effectiveness through experimental results on benchmark problems. Yu and Xu (2015) investigate the problem of parameter uncertainty and how it can be addressed within the robust MDP framework. They propose a distributionally robust counterpart formulation that allows for a more robust decision-making process in MDP, especially when the probability distribution of the uncertain parameters is not precisely known. The authors also present theoretical results for the distributionally robust counterpart formulation, including the existence of an optimal policy and the convergence of value iteration algorithms. The authors further demonstrate the effectiveness of their approach through numerical experiments on various MDP problems. Taking a different approach, Lim and Autef (2019) propose a kernel-based reinforcement learning approach to solve robust MDP. They aim to address the challenges posed by uncertainties in the transition probabilities. Specifically, they applied a kernel-based method to estimate

the transition probabilities and incorporated a robust optimization technique to minimize the expected costs under the worst-case scenario, taking into account the uncertainties in the transition probabilities. In a similar vein, Abdullah et al. (2019) propose a novel approach to address robustness issues in MDP, that is, a Wasserstein distance-based robust optimization framework that can effectively handle parameter uncertainty and minimize the worst-case cost. Specifically, the proposed algorithm utilizes a Wasserstein metric to quantify the distance between the nominal and uncertain distributions of transition probabilities in MDP. Building on these advancements, Y. Wang and Zou (2021, 2022) propose an online learning algorithm and policy gradient method that incorporates model uncertainty in robust MDP. This method is designed to minimize the cost function under the worst-case scenario while considering the possibility of inaccurate parameter estimates. The authors demonstrate that their method outperforms existing methods in terms of both performance and robustness on various benchmark problems.

In summary, risk-sensitive MDP and robust MDP are distinct approaches for decision-making in uncertain environments. They vary in terms of their objectives, uncertainty modeling techniques, and optimal solutions. Risk-sensitive MDP aims to minimize risk while balancing expected rewards, using risk measures such as variance or CVaR. The objective is to find a policy that achieves a good balance between expected rewards and risk. On the other hand, robust MDP aims to ensure robust performance in uncertain environments by finding a policy that performs well across a range of possible models or scenarios. Both approaches address decision-marking under uncertainty but emphasize different aspects: Risk-sensitive MDP focuses on the trade-off between expected reward and risk, while robust MDP focuses on maintaining robust performance in uncertain environments.

4.3 Robust MDP with Hidden Regime Rules

The introduction of hidden regime rules in a Markov Decision Process (MDP) framework creates a unique and challenging problem. The underlying premise of this approach is to account for the inherent uncertainties in the transition probabilities, which are contingent on the current state and the financial market's hidden regime. However, solving such a robust

MDP can be quite complex due to the need to effectively handle these uncertainties and still generate optimal strategies. Dynamic programming serves as an effective and efficient approach to addressing this complexity. It aids in determining the optimal action at each stage based on the current state and hidden regime, enhancing the policy's robustness in response to market dynamics. In this section, we will discuss the robust DP framework for solving the robust MDP with hidden regime rules. For the convenience of readers, we provide Table 4.1 that includes all symbols in this section.

Table 4.1: Notations for the Robust MDP with Hidden Regime Rules

Symbols	**Explanations for the Notation**
$\mathscr{S}_t, \mathscr{A}_t, \mathscr{I}_t, \mathscr{H}_t$	the space of state, action, hidden regime, and history at time t
s_t, a_t, i_t, h_t	state, action, hidden regime, and history at time t
ϖ_t, ν_t	history-dependent hidden regime rule and decision rule at time t
$\mathscr{T}^{\varpi}$	the set of all conditional measure consistent with a Markov hidden regime rule
ϕ, ψ	the sequence of hidden regime rules and decision rules
$\mathscr{T}^{\phi}$	the set of all conditional measure consistent with the regime change ϕ
ψ^{ϕ}	the sequence of regime-changing dependent policies
$V^{\psi^{\phi}}(\cdot)$	the value function of regime-changing dependent policies
$V_t^*(\cdot)$	the optimal value function at time t
ϖ	the set of all history dependent policies with regime changes
ϖ_n	the set of all history dependent randomized policies with regime changes for epoch $t \geq n$
$\mathscr{D}_n$	the set of all history dependent decision rules that incorporate with the hidden regime rule ϖ_n at epoch n
ϖ_D	the set of all history-dependent deterministic policies with regime changes
ϖ_{MD}	the set of all regime-changing dependent Markov deterministic policies
$\mathscr{D}^{\varpi}$	the subset of all Markov deterministic decision rules that incorporate the stationary hidden regime rule ϖ
$\mathscr{L}_{\mathscr{D}^{\varpi}}$	the robust Bellman operator of decision rules that are dependent on the stationary hidden regime rule ϖ
$\mathscr{P}_{\mathscr{E}}$	the entropy uncertainty set over transition probabilities

4.3.1 Finite Horizon Robust DP

In a finite-time horizon $t \in T = \{0, 1, ..., N-1\}$ for some $N \geq 1$, decision epochs refer to discrete points where decisions are made. We consider an MDP with a finite state space $\mathscr{S}_t$ and a finite action space $\mathscr{A}_t$ for all $t \in T$. At each timestep $t \in T$, the system is assumed to occupy a finite state $s_t \in \mathscr{S}_t$, where $\mathscr{S}_t$ is assumed to be discrete or finite state space. Let $\mathscr{H}_t$ denote the set of all histories h_t, where the history h_t contains the historical states up to time t ($h_t = (s_0, s_1, ..., s_t)$). For any discrete set $\mathscr{B}$, we denote by $\mathscr{M}(\mathscr{B})$ the set of probability measure on $\mathscr{B}$. The hidden regime i_t represents different underlying states of the market which aren't directly observable but have a substantial influence on observable market behaviors. Examples of such regimes include bull markets, bear markets, and stable markets. We denote by $\mathscr{I}_t$ the hidden regime space representing a set of all possible specific hidden regimes i_t that the underlying system can be in at time t. We define ϖ_t as history-dependent hidden regime rules at time t, where $\varpi_t : \mathscr{H}_t \to \mathscr{M}(\mathscr{I}_t)$ represents the probability distribution over the finite hidden regimes space $\mathscr{I}_t$ based on the history of states up to time t. Decision makers can choose actions either randomly or deterministically based on the hidden regime rules and current state. A random action corresponds to a state $s_t \in \mathscr{S}_t$ and a hidden regime rule $\varpi_t \in \mathscr{M}(\mathscr{I}_t)$, which corresponds to an element $q_{s_t,\varpi_t} \in \mathscr{M}(\mathscr{A}_t)$. In this context, an action $a \in \mathscr{A}_t$ is selected with probability $q_{s_t,\varpi_t}(a)$. At each timestep $t \in T$, a state $s_t \in \mathscr{S}_t$ and hidden regime rule $\varpi_t \in \mathscr{M}(\mathscr{I}_t)$ consist a set of conditional measures $\mathscr{P}_t(s_t, \varpi_t) \subseteq \mathscr{M}(\mathscr{S}_{t+1})$ with the interpretation that under the hidden regime rules ϖ_t, then the next state s_{t+1} is derived from the conditional measure $p_{s_t,\varpi_t} \in \mathscr{P}_t(s_t, \varpi_t)$. These conditional measures play a crucial role as they encapsulate the uncertainty in the transition probability of the Markov Decision Processes (MDPs). This uncertainty arises from the underlying state transitions and regime change governed by environment change.

The set of all conditional measures consistent with a history dependent hidden regime rule ϖ_t is given by

$$\mathscr{T}^{\varpi_t} = \Big\{ p : \mathscr{H}_t \to \mathscr{M}(\mathscr{S}_{t+1}) : \forall h \in \mathscr{H}_t, s \in \mathscr{S}_{t+1}, p_h(s) = p_{s_t,\varpi_t(h)}(s), \\ p_{s_t,\varpi_t(h)} \in \mathscr{P}_t(s_t, \varpi_t(h)) \Big\}. \tag{4.1}$$

A hidden regime rule ϖ_t is called Markovian if it is a function of the current state s_t alone.

The set of all conditional measure consistent with a Markov hidden regime rule $\varpi_t(s_t)$ is given by

$$\mathscr{T}^{\varpi_t} = \Big\{p : \mathscr{S}_t \to \mathscr{M}(\mathscr{S}_{t+1}) : \forall s \in \mathscr{S}_t, p_s \in \mathscr{P}_t(s, \varpi_t(s))\Big\},$$

i.e., for every state $s \in \mathscr{S}_t$, the next state can be determined by any $p \in \mathscr{P}_t(s, \varpi_t(s))$.

A decision rule ν_t is a procedure for choosing actions at a specified decision epoch $t \in T$ dependent on the history of the states, which is a mapping $\nu_t : \mathscr{H}_t \to \mathscr{M}(\mathscr{A}_t)$. A decision rule ν_t is called deterministic if probability measures that assign all the probability mass to a single action $a \in \mathscr{A}_t$, and Markovian if it only depends on the current state s_t instead of the entire history up to time t. Let's consider a policy ψ as a sequence of decision rules denoted by $\psi = \Big(\nu_t; t \in T\Big)$, where each ν_t represents the decision rule at time t. In a similar vein, a regime change ϕ is defined as a sequence of hidden regime rules denoted by $\phi = \Big(\varpi_t; t \in T\Big)$. Given a hidden regime rule ϖ_t at time t, we can define a hidden regime rule dependent decision rule as $\nu_t(\cdot|\varpi_t)$. This decision rule takes into account the hidden regime rule ϖ_t to determine the optimal action to be taken. To incorporate the regime changes, we introduce the concept of the regime-changing dependent policy, denoted as ψ^ϕ. This policy utilizes decision rules derived from underlying hidden regime rules, represented as $\psi^\phi = \Big(\nu_t(\cdot|\varpi_t), t \in T\Big)$. In other words, the decision rules in ψ^ϕ are applied at each decision epoch based on the respective hidden regime rules. A regime change ϕ induces a collection of measures in the history space $\mathscr{H}_N$ due to the uncertainty of the conditional measures. Consequently, we assume that the set $\mathscr{T}^\phi$ of measures consistent with ϕ possesses the following structure.

Assumption 4.3.1 *The set $\mathscr{T}^\phi$ of measures consistent with the regime change ϕ is given by*

$$\begin{aligned}\mathscr{T}^\phi &= \Big\{p : \forall h_N \in \mathscr{H}_N, p(h_N) = \prod_{t\in T} p_{h_t}(s_{t+1}), p_{h_t} \in \mathscr{T}^{\varpi_t}, t \in T\Big\} \\ &= \mathscr{T}^{\varpi_0} \times \mathscr{T}^{\varpi_1} \times \mathscr{T}^{\varpi_2} ... \times \mathscr{T}^{\varpi_{N-1}}.\end{aligned}$$

The decision maker receives a reward $r_t(s_t, a_t, s_{t+1})$ when the action $a_t \in \mathscr{A}_t$ is chosen in state $s_t \in \mathscr{S}_t$ at the decision epoch t, and the state at the next epoch is $s_{t+1} \in \mathscr{S}_{t+1}$. Since s_{t+1} is ambiguous, we allow the reward at time t to depend on s_{t+1} as well. Without loss of generality, we can assume that the reward is certain. The reward function $r_N(s)$ at the epoch N is only a function of the state $s \in \mathscr{S}_N$.

The reward $V_0^{\psi^\phi}(s)$ is generated by a regime-changing dependent policy starting from the initial state $s_0 = s$, which is defined as follows.

$$V_0^{\psi^\phi}(s) = \inf_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big], \tag{4.2}$$

where $\mathbb{E}^p$ denotes the expectation with respect to the fixed measure $p \in \mathscr{T}^\phi$. Equation (4.2) defines the value function of regime-changing policy ψ^ϕ to be the minimum expected rewards, considering all measures that are consistent with the regime change ϕ. This approach is commonly known as the robust approach in the conventional optimization literature (Ben-Tal & Nemirovski, 1998). It ensures that the policy performs well under various possible scenarios and provides a strong foundation for decision-making. Let Π denotes the set of all history dependent policies with regime changes. Then, the goal of robust DP aims to seek the optimal robust value function

$$V_0^*(s) = \sup_{\psi^\phi \in \Pi} \left\{ \inf_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \right\}. \tag{4.3}$$

For understanding the implications of the rectangularity assumption the objective (4.3) has to be interpreted in an adversarial setting: The decision maker chooses a regime-changing dependent policy ψ^ϕ; an adversary observes the policy ψ^ϕ and selects a probability measure $p \in \mathscr{T}^\phi$ that minimizes the reward. In this framework, rectangularity serves as an independence assumption, implying that the selection of a particular distribution $\bar{p} \in \mathscr{P}(s_t, \varpi_t)$ at time t does not impose any constraints on the future choices of the adversary. This leads to a separability that is crucial for establishing the robust counterpart of the Bellman recursive function. It is worth noting that the rectangularity assumption may not always be appropriate because the shape of the state space may not be rectangular. However, in certain situations, by considering the time-inhomogeneity, we can adopt the rectangularity assumption to simplify the problem.

The optimistic value $\bar{V}_0^{\psi^\phi}(s)$ of a policy ψ^ϕ starting from the initial state $s_0 = s$ is defined as

$$\bar{V}_0^{\psi^\phi}(s) = \sup_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big].$$

Analogous to the robust value function V_0^*, the optimistic value function $\bar{V}_0^*$ is defined as

$$\bar{V}_0^* = \sup_{\psi^\phi \in \Pi} \Big\{ \bar{V}_0^{\psi^\phi}(s) \Big\} = \sup_{\psi^\phi \in \Pi} \Big\{ \sup_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \Big\}.$$

The regime-changing dependent value function, denoted as $V_n^{\psi^\phi}(h_n)$, is obtained by using a regime-changing dependent policy ψ^ϕ over epochs $n, n+1, ..., N-1$, starting from the history h_n. It is defined as:

$$V_n^{\psi^\phi}(h_n) = \inf_{p \in \mathscr{T}_n^\phi} \mathbb{E}^p \Big[\sum_{t=n}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big]. \tag{4.4}$$

The set of conditional measures $\mathscr{T}_n^\phi$ consistent with the regime change ϕ and the history h_n is given by:

$$\begin{aligned} \mathscr{T}_n^\phi &= \Big\{ p_n : \mathscr{H}_n \to \prod_{t=n}^{N-1} \mathscr{M}(\mathscr{S}_{t+1}) : \forall h_n \in \mathscr{H}_n, p_{h_n}(s_{n+1}, ..., s_N) = \prod_{t=n}^{N-1} p_{h_t}(s_{t+1}), p_{h_t} \in \mathscr{T}_{t=n,...,N-1}^{\varpi_t} \Big\} \\ &= \mathscr{T}^{\varpi_n} \times \mathscr{T}^{\varpi_{n+1}} \times ... \times \mathscr{T}^{\varpi_{N-1}} = \mathscr{T}^{\varpi_n} \times \mathscr{T}_{n+1}^\phi, \end{aligned} \tag{4.5}$$

where $\mathscr{T}^{\varpi_n}$ is the set of conditional measures for policy ϖ_n and $\mathscr{T}_{n+1}^\phi$ is the set of conditional measures consistent with the regime change ϕ from epochs $n+1$ to N.

The optimal value function $V_n^*(h_n)$ is defined as the supremum of the value function $V_n^{\psi^\phi}(h_n)$ over all policies ψ^ϕ starting from the history h_n at epoch n. Now, let $V_n^*(h_n)$ be the optimal value function, it can be derived as follows:

$$V_n^*(h_n) = \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{p \in \mathscr{T}_n^\phi} \mathbb{E}^p \Big[\sum_{t=n}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \right\}, \tag{4.6}$$

where ϖ_n represents the set of all history-dependent randomized policies with regime changes for epoch $t \geq n$. The following theorem establishes the robust counterpart of the robust Bellman recursive function.

Theorem 4.3.2 *The set of value functions* $\{V_n^* : n = 0, 1, 2, ..., N\}$ *satisfies the following robust Bellman equation:*

$$\begin{aligned} V_N^*(h_N) &= r_N(s_N), \\ V_n^*(h_n) &= \sup_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(h_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}^*(h_n, s') \Big] \right\}, \quad n = 0, 1, 2, \ldots, N-1. \end{aligned} \tag{4.7}$$

Proof. By equation (4.5) and (4.6), we can derive the following

$$V_n^*(h_n) = \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{p=(\bar{p},P)\in \mathscr{T}^{\varpi_n} \times \mathscr{T}_{n+1}^{\phi}} \mathbb{E}^p \Big[\sum_{t=n}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \right\}. \tag{4.8}$$

Since the conditional measures P does not affect the first term $r_n(s_n, \nu_t(h_n|\varpi_n), s_{n+1})$, we have

$$\begin{aligned}
V_n^*(h_n) &= \sup_{\psi^\phi \in \varpi_n} \left\{ \inf_{p\in \mathscr{T}_n^{\phi}} \mathbb{E}^p \Big[\sum_{t=n}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \right\} \\
&= \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{(\bar{p},P)\in \mathscr{T}^{\varpi_n} \times \mathscr{T}_{n+1}^{\phi}} \mathbb{E}^{\bar{p}} \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{t+1}) + \mathbb{E}^P \Big[\sum_{t=n+1}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \Big] \right\} \\
&= \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{\bar{p}\in \mathscr{T}^{\varpi_n}} \mathbb{E}^{\bar{p}} \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + \inf_{P\in \mathscr{T}_{n+1}^{\phi}} \mathbb{E}^P \Big[\sum_{t=n+1}^{N-1} r_t(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) + r_N(s_N) \Big] \Big] \right\} \\
&= \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^{\psi^\phi}(h_n, s_{n+1}) \Big] \right\},
\end{aligned} \tag{4.9}$$

where the last equality follows from the definition of $V_{n+1}^{\psi^\phi}(h_{n+1})$.

Therefore, equation (4.9) implies that

$$\begin{aligned}
V^*(h_n) &\leq \sup_{\psi^\phi \in \varpi_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^*(h_n, s_{n+1}) \Big] \right\} \\
&= \sup_{\nu_n(h_n|\varpi_n)\in \mathscr{D}_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^*(h_n, s_{n+1}) \Big] \right\},
\end{aligned} \tag{4.10}$$

where $\mathscr{D}_n$ is the set of all history-dependent decision rules that incorporate the hidden regime rule ϖ_n at epoch n.

Since $V_{n+1}^*(h_{n+1}) = \sup_{\psi^\phi \in \Pi_{n+1}} \{V_{n+1}^{\varpi}(h_{n+1})\}$, it follows that for all, $\epsilon > 0$ there exists a policy $\psi_{n+1}^\epsilon \in \Pi_{n+1}$ such that $V_{n+1}^{\psi_{n+1}^\epsilon}(h_{n+1}) \geq V_{n+1}^*(h_{n+1}) - \epsilon$, for all $h_{n+1} \in \mathscr{H}_{n+1}$. For all $\nu_n(h_n|\varpi_n) \in \mathscr{D}_n$, $(\nu_n(h_n|\varpi_n), \psi_{n+1}^\epsilon) \in \Pi_n$. Therefore,

$$\begin{aligned}
V^*(h_n) &= \sup_{\psi^\phi \in \Pi_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^{\psi^\phi}(h_n, s_{n+1}) \Big] \right\} \\
&\geq \sup_{\nu_n(h_n|\varpi_n)\in \mathscr{D}_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^{\varpi_{n+1}^\epsilon}(h_n, s_{n+1}) \Big] \right\} \\
&\geq \sup_{\nu_n(h_n|\varpi_n)\in \mathscr{D}_n} \left\{ \inf_{p\in \mathscr{T}^{\varpi_n}} \mathbb{E}^p \Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V_{n+1}^*(h_n, s_{n+1}) \Big] \right\} - \epsilon.
\end{aligned} \tag{4.11}$$

Since $\epsilon > 0$ is arbitrary, by combining (4.10) and (4.11), we have

$$V^*(h_n) = \sup_{\nu_n(h_n|\varpi_n)\in\mathscr{D}_n} \left\{ \inf_{p\in\mathscr{T}^{\varpi_n}} \mathbb{E}^p\Big[r_n(s_n, \nu_n(h_n|\varpi_n), s_{n+1}) + V^*_{n+1}(h_n, s_{n+1})\Big]\right\}. \tag{4.12}$$

By definition of (4.1), we have

$$\begin{aligned}
V^*_n(h_n) &= \sup_{a\in\mathscr{A}} \inf_{p_{s_n,\varpi_n(h_n)}\in\mathscr{P}(s_n,\varpi_n(h_n))} \left\{ \sum_{s'\in\mathscr{S}_{n+1}} p_{s_n,\varpi_n(h_n)}(s')[r_n(s_n,a,s') + V^*_{n+1}(h_n,s')]\right\} \\
&= \sup_{a\in\mathscr{A}} \left\{ \inf_{p_{s_n,\varpi_n(h_n)}\in\mathscr{P}(s_n,\varpi_n(h_n))} \Big[\sum_{s'\in\mathscr{S}_{n+1}} p(s')[r_n(s_n,a,s') + V^*_{n+1}(h_n,s')]\Big]\right\} \\
&= \sup_{a\in\mathscr{A}} \left\{ \inf_{p\in\mathscr{P}(s_n,\varpi_n(h_n))} \Big[\sum_{s'\in\mathscr{S}_{n+1}} p(s')[r_n(s_n,a,s') + V^*_{n+1}(h_n,s')]\Big]\right\} \\
&= \sup_{a\in\mathscr{A}} \left\{ \inf_{p\in\mathscr{P}(s_n,\varpi_n(h_n))} \mathbb{E}^p\Big[r_n(s_n,a,s') + V^*_{n+1}(h_n,s')\Big]\right\}.
\end{aligned} \tag{4.13}$$

This completes the proof. ■

The following corollary demonstrates that the decision maker can achieve the same robust reward by considering only deterministic policies, without the need for considering randomized policies.

Corollary 4.3.1 *Let Π_D be the set of all history-dependent deterministic policies with regime changes. Then, Π_D is adequate for characterizing the value functions V_n in the sense that for all $n = 0, ..., N-1$.*

$$V^*_n(h_n) = \sup_{\psi^\phi\in\varpi_D} \Big\{ V^{\psi^\phi}_n(h_n)\Big\}. \tag{4.14}$$

Proof. From Theorem 4.7, we can drive the following:

$$\begin{aligned}
V^*_n(h_n) &= \sup_{a\in\mathscr{A}} \left\{ \inf_{p\in\mathscr{P}(s_n,\varpi_n(h_n))} \mathbb{E}^p\Big[r_n(s_n,a,s') + V^*_{n+1}(h_n,s')\Big]\right\} \\
&= \sup_{\psi^\phi\in\Pi_D} \left\{ \inf_{p\in\mathscr{P}(s_n,\varpi_n(h_n))} \mathbb{E}^p\Big[r_n(s_n,\nu_n(h_n|\varpi_n),s') \right. \\
&\qquad \left. + \inf_{\bar{P}\in\mathscr{T}^\phi_{n+1}} \mathbb{E}^{\bar{P}}\Big[\sum_{t=n+1}^{N-1} r_t(s_t,\nu_t(h_t|\varpi_t),s') + r_N(s_N)\Big]\right\} \\
&= \sup_{\psi^\phi\in\Pi_D} \left\{ \inf_{p\in\mathscr{T}^\phi_n} \mathbb{E}^p\Big[\sum_{t=n}^{N-1} r_t(s_t,\nu_t(h_t|\varpi_t),s_{t+1}) + r_N(s_N)\Big]\right\} \\
&= \sup_{\psi^\phi\in\Pi_D} \Big\{ V^{\psi^\phi}_n(h_n)\Big\}.
\end{aligned}$$

This completes the proof. ■

Next, we establish that the decision maker can limit themselves to deterministic Markov policies, which are policies that the decision rule at any given epoch is solely based on the current state s_n and not on the history h_n.

Theorem 4.3.3 *(Markov Optimality) For all $n = 0, 1, ..., N$, the robust value function $V_n^*(h_n)$ is a function of the current state s_n alone and $V_n^*(s_n) = \sup_{\psi^\phi \in \Pi_{MD}} \left\{V_n^{\psi^\phi}(s_n)\right\}$, where Π_t is Markov hidden regime rules, and ϖ_{MD} is the set of all regime-changing dependent deterministic Markov policies. Therefore, the robust Bellman equation can be derived as follows:*

$$V_N^*(s_N) = r_N(s_N),$$

$$V_n^*(s_n) = \sup_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(s_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}^*(s') \Big] \right\}, \forall n \in T. \tag{4.15}$$

Proof. The result is established by induction on the epoch t for all $t > n$. For $t = N$, the value function $V_N^*(h_N) = r_N(s_N)$ and it is a function of only the current state. Starting with the Bellman equation (4.7), we have

$$V^*(h_n) = \sup_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(h_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}^*(h_n, s') \Big] \right\}. \tag{4.16}$$

Since $\varpi_n(\cdot)$ is Markovian function dependent only on current state, therefore, the equation (4.16) can be rewritten as follows

$$V_n^*(h_n) = \sup_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(s_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}^*(h_n, s') \Big] \right\}. \tag{4.17}$$

Since the right-hand side of equation (4.17) is solely dependent on h_n through s_n, we can further simplify it to:

$$V_n^*(s_n) = \sup_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(s_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}^*(s') \Big] \right\}. \tag{4.18}$$

This completes the proof. ■

The recursive relation of equation (4.18) is the foundation for conventional robust DP problem. This relation establishes that the optimal value function $V_n^*(s)$ can be obtained by maximizing the inner expression over all possible actions $a \in \mathscr{A}$, and taking the infimum

over the conditional measure $p \in \mathscr{P}(s_n, \varpi_n(s_n))$ for next epoch. Suppose the action set $\mathscr{A}$ is finite. Then the optimal decision rule ν_n^* at epoch n is given by

$$\nu_n^*(s) = \arg\max_{a \in \mathscr{A}} \left\{ \inf_{p \in \mathscr{P}(s_n, \varpi_n(s_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + V_{n+1}(s') \Big] \right\}.$$

Hence, efficient computation of the value function V_n^* requires the ability to solve the inner optimization problem efficiently. It is worth noting that the Theorem 4.3.3 implies the following result for the optimistic value function $\bar{V}_n^*$.

Theorem 4.3.4 *For $n = 0, ..., N$, the optimistic value function $\bar{V}_n^*(h_n)$ is a function of the current state s_n alone, and it can be derived as follows:*

$$\bar{V}_n^*(s_n) = \sup_{\psi^\phi \in \Pi_{MD}} \Big\{ \bar{V}_n^{\psi^\phi}(s_n) \Big\}, n \in T,$$

where ϖ_{MD} is the set of all regime-changing deterministic Markov policies. Therefore,

$$\bar{V}_n^*(s_n) = \sup_{a \in \mathscr{A}} \left\{ \sup_{p \in \mathscr{P}(s_n, \varpi_n(s_n))} \mathbb{E}^p \Big[r_n(s_n, a, s') + \bar{V}_{n+1}^*(s') \Big] \right\}, \forall n \in T.$$

4.3.2 Infinite Horizon Robust DP

In this section, we address an infinite horizon robust DP with regime-changing dependent policy. The settings for the infinite horizon case is similar to the finite case, where the system contains the state $s \in \mathscr{S}$. The state space $\mathscr{S}$ is assumed to be finite and discrete, the decision maker is allowed to choose an action $a \in \mathscr{A}$ from a finite or discrete action set. The main difference is that time t belongs to the set of $t \in T = [0, 1, ...]$. We assume that the reward function $r(s_t, a_t, s_{t+1})$, depends on the current state s_t, the chosen action a_t, and next state s_{t+1}. Furthermore, it is also bounded such that $\sup_{s \in \mathscr{S}, a \in \mathscr{A}} \{r(s, a, s')\} = \mathcal{R} < \infty$. The value function $V^{\psi^\phi}(s)$ represents the expected cumulative discounted reward under a regime-changing dependent policy ψ^ϕ when the initial state $s_0 = s$, and it is defined as follows:

$$V^{\psi^\phi}(s) = \inf_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, \nu_t(h_t | \varpi_t), s_{t+1}) \Big],$$

where $\mathscr{T}^\phi = \prod_{t \in T} \mathscr{T}^{\varpi_t}$, and $\gamma \in (0, 1)$ is a discount factor. It is clear that all regime-changing dependent policies ψ^ϕ, $\sup_{s \in \mathscr{S}} \{V^{\psi^\phi}(s)\} \leq \mathcal{R}/(1-\gamma)$. The optimal value function

in state s is given by

$$V^*(s) = \sup_{\psi^\phi \in \Pi} \{V^{\psi^\phi}(s)\} = \sup_{\psi^\phi \in \Pi} \Big\{ \inf_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, \nu_t(h_t|\varpi_t), s_{t+1}) \Big] \Big\},$$

where Π is the set of all regime-changing dependent policies that take into account the entire history of states. It can be derived similarly as in previous subsection that if we restrict the decision maker to deterministic Markov policies and assume that the regime change ϕ is a Markov function of the current state, then the optimal value function $V^*(s)$ remains the same without losing the performance. In other words, the decision maker can achieve the same performance by restricting the policy to deterministic Markov policies, which can be expressed as follows:

$$V^*(s) = \sup_{\psi^\phi \in \Pi_{\text{MD}}} \Big\{ V^{\psi^\phi}(s) \Big\}.$$

Let $\mathbf{V}$ denote the set of all bounded real-valued functions on the discrete set on $\mathscr{S}$. The L_∞ norm on $\mathbf{V}$ is denoted as $||V||$, given by

$$||V|| = \max_{s \in \mathscr{S}} |V(s)|.$$

Therefore, $(\mathbf{V}, ||\cdot||)$ is a Banach space. Let $\mathscr{D}^\varpi$ be any subset of all Markov deterministic decision rules that is dependent on the stationary and Markov hidden regime rule ϖ. Then, we define the robust Bellman operator $\mathscr{L}_{\mathscr{D}^\varpi}$ on $\mathbf{V}$ as follows:
For all $V \in \mathbf{V}$,

$$\mathscr{L}_{\mathscr{D}^\varpi} V(s) = \sup_{\nu(s|\varpi) \in \mathscr{D}^\varpi} \Big\{ \inf_{p \in \mathscr{P}(s, \varpi(s))} \mathbb{E}^p \Big[r(s, \nu(s|\varpi), s') + \gamma V(s') \Big] \Big\}, \forall s, s' \in \mathscr{S}.$$

Theorem 4.3.5 *(Bellman Equation). The operator $\mathscr{L}_{\mathscr{D}^\varpi}$ satisfies the following properties:*

1. *The operator $\mathscr{L}_{\mathscr{D}^\varpi}$ is a contraction mapping on $\mathbf{V}$; in particular for all $U, V \in \mathbf{V}$,*

$$||\mathscr{L}U - \mathscr{L}V|| \leq \gamma ||U - V||.$$

2. *The operator equation $\mathscr{L}_{\mathscr{D}^\varpi} V = V$ has a unique solution. Moreover,*

$$V(s) = \sup_{\nu(s|\varpi) \in \mathscr{D}^\varpi} \inf_{p \in \mathscr{T}^\phi} \mathbb{E}^p \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, \nu_t(s_t|\varpi_t), s_{t+1}) \Big].$$

Proof. Let $U, V \in \mathbf{V}$. Fix $s \in \mathscr{S}$ and assume that $\mathscr{L}U(s) \geq \mathscr{L}V(s)$. Fix $\epsilon > 0$, and choose $\nu(\cdot|\varpi) \in \mathscr{D}^{\varpi}$ such that for all $s \in \mathscr{S}$,

$$\inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^{p}[r(s, \nu(s|\varpi), s') + \gamma U(s')] \geq \mathscr{L}_{\mathscr{D}^{\varpi}} U(s) - \epsilon.$$

Choose a conditional probability measure $p_s \in \mathscr{P}(s, \varpi(s)), s \in \mathscr{S}$ such that

$$\mathbb{E}^{p_s}[r(s, \nu(s|\varpi), s') + \gamma V(s')] \leq \inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^{p}[r(s, \nu(s|\varpi), s') + \gamma V(s')] + \epsilon.$$

Then

$$\begin{aligned}
0 &\leq \mathscr{L}U(s) - \mathscr{L}V(s) \\
&\leq \Big(\inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^{p}[r(s, \nu(s|\varpi), s') + \gamma U(s')] + \epsilon \Big) - \Big(\inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^{p}[r(s, \nu(s|\varpi), s') + \gamma V(s')] \Big) \\
&\leq \Big(\mathbb{E}^{p_s}[r(s, \nu(s|\varpi), s') + \gamma U(s')] + \epsilon \Big) - \Big(\mathbb{E}^{p_s}[r(s, \nu(s|\varpi), s') + \gamma V(s')] - \epsilon \Big) \\
&\leq \gamma \mathbb{E}^{p_s}[U - V] + 2\epsilon \\
&\leq \gamma ||U - V|| + 2\epsilon.
\end{aligned}$$

Repeating the argument for the case $\mathscr{L}U(s) \leq \mathscr{L}V(s)$ implies that

$$|\mathscr{L}U(s) - \mathscr{L}V(s)| \leq \gamma ||U - V|| + 2\epsilon.$$

Since $\mathscr{L}_{\mathscr{D}^{\varpi}}$ is a contraction operator on a Banach space, the Banach fixed point theorem implies that the operator equation $\mathscr{L}_{\mathscr{D}^{\varpi}} V = V$ has a unique solution for $V \in \mathbf{V}$. Fix the hidden regime rules ϖ such that $\nu_t(\cdot|\varpi) \in \mathscr{D}$, for all $t \geq 0$. Then, for all $n \geq 0$,

$$\begin{aligned}
V(s) = \mathscr{L}_{\mathscr{D}^{\varpi}} V(s) &\geq \inf_{p \in \mathscr{T}^{\phi}} \mathbb{E}^{p}\Big[\sum_{t=0}^{n} r(s_t, \nu_t(s_t|\varpi), s_{t+1}) + \gamma^{n+1} V(s_{n+1}) \Big] \\
&= \inf_{p \in \mathscr{T}^{\phi}} \mathbb{E}^{p}\Big[\sum_{t=0}^{\infty} r(s_t, \nu_t(s_t|\varpi), s_{t+1}) + \gamma^{n+1} V(s_{n+1}) - \sum_{t=n+1}^{\infty} r(s_t, \nu_t(s_t|\varpi), s_{t+1}) \Big] \\
&\geq V^{\psi^{\phi}}(s) - \gamma^{n+1} ||V|| - \frac{\gamma^{n+1} \mathcal{R}}{1 - \gamma},
\end{aligned} \tag{4.19}$$

where $||V|| = \max_{s \in \mathscr{S}} ||V(s)||$. Since n is arbitrary, it follows that $V(s) \geq \sup_{\nu_t(s|\varpi) \in \mathscr{D}^{\varpi}} \big\{ V^{\psi^{\phi}}(s) \big\}$.

Fix $\epsilon > 0$ and hidden regime rule ϖ, choosing a hidden regime dependent deterministic decision rule $\nu(s|\varpi) \in \mathscr{D}^{\varpi}$ such that for all $s \in \mathscr{S}$,

$$V(s) = \mathscr{L}_{\mathscr{D}^{\varpi}} V(s) \leq \inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^{p}\Big[r(s, \nu(s|\varpi), s') + \gamma V(s') \Big] + \epsilon. \tag{4.20}$$

Then, we have

$$V(s) \leq V^{\psi^\phi}(s) + \gamma^n ||V|| + \frac{\epsilon}{1-\gamma}, \quad \forall n \geq 0.$$

Since ϵ and n are arbitrary, it follows equation (4.19) and (4.3.2) that

$$V(s) = \sup_{\nu(s|\varpi) \in \mathscr{D}^\varpi} \Big\{ V^{\varpi^\phi}(s) \Big\}.$$

This completes the proof. ■

Corollary 4.3.2 *The properties of the operator $\mathscr{L}_{\mathscr{D}^\varpi}$ implies the following:*

1. *Let $\nu(\cdot|\varpi)$ be any deterministic Markov decision rules incorporated with a hidden regime rule. Then, the value function V^{ψ^ϕ} of the stationary regime-changing dependent policy $\Big(\nu(\cdot|\varpi)\Big)$ is the unique solution of the Bellman operator equation*

$$V(s) = \inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^p \Big[r(s, \nu(s|\varpi), s') + \gamma V(s') \Big], s \in \mathscr{S},$$

   where ϖ represents a stationary and Markov hidden regime rule.

2. *The value function V^* is the unique solution of the operator equation*

$$V(s) = \sup_{a \in \mathscr{A}} \inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^p \Big[r(s, a, s') + \gamma V(s') \Big], s \in \mathscr{S}.$$

Moreover, for all $\epsilon > 0$, there exists an ϵ-optimal stationary policy; i.e., there exists $\psi^\epsilon = (\nu^\epsilon, \nu^\epsilon, ...)$ such that $V^{\varpi^\epsilon} \geq V^* - \epsilon$.

Proof. The results follow Corollary 3.1 in Iyengar (2005) by setting $\mathscr{D}^\varpi = \{\nu(\cdot|\varpi)\}$ and $\mathscr{D}^\varpi = \prod_{s \in \mathscr{S}} \mathscr{A}(s)$, respectively. This completes the proof. ■

Next, we will show an example of Entropy model that model the uncertainty over the transition probability. Then, we address the inner problem with uncertainty on the transition probability by showing that the inner problem is equivalent to the dual problem, and the dual problem can be approximated by Bisection Algorithm.

4.3.3 An Example: Entropy Uncertainty Set

Prior subsections have been committed to the expansion of robust dynamic programming (DP) results that incorporate hidden regime rules. Now, our attention is pivoted to tackling the inner problem represented by:

$$\inf_{p \in \mathscr{P}(s,\varpi(s))} \mathbb{E}^p[V]$$

in which the challenge is to address the uncertainty inherent in transition probabilities. In this subsection, we present an entropy model that captures the inherent uncertainty in transition probabilities. The transition probabilities in an MDP represent the likelihood of transitioning from one state to another. However, in many real-world scenarios, these transition probabilities may not be precisely known. The entropy model provides a probabilistic framework that considers a set of possible transition probability distributions to account for this uncertainty. The set of distributions is characterized by a constraint on the Kullback-Leibler (KL) divergence between a reference distribution and the uncertain distribution. The reference distribution, denoted as q, represents our prior knowledge or belief about the transition probabilities, while the uncertain distribution, denoted as p, represents the true but unknown distribution. The KL divergence measures the difference between two probability distributions.

The objective of the entropy uncertainty is to find a distribution $p_{s,\varpi(s)}$ associated with hidden regime rule ϖ within the set $\mathscr{P}(s, \varpi(s))$ that satisfies the constraint:

$$D(p_{s,\varpi(s)}||q_{s,\varpi(s)}) \leq \beta,$$

where $D(p_{s,\varpi(s)}||q_{s,\varpi(s)})$ is the KL divergence between hidden regime dependent distributions $p_{s,\varpi(s)}$ and $q_{s,\varpi(s)}$, and β is a fixed value representing an upper bound on the allowed divergence. Note that $q_{s,\varpi(s)}$ is the reference distribution, which can be estimated from empirical data or expert knowledge. The condition $\beta > 0$, along with $q_{s,\varpi(s)} > 0$, ensures that the set $\mathscr{P}(s, \varpi(s))$ has a nonempty interior. The Kullback-Leibler (KL) divergence can be formulated as:

$$D(p_{s,\varpi(s)}||q_{s,\varpi(s)}) = \sum_{s \in \mathscr{S}} p_{s,\varpi(s)} \log \frac{p_{s,\varpi(s)}}{q_{s,\varpi(s)}}.$$

Solving the MDP problem with the entropy model involves finding a policy that maximizes the expected total reward while considering the set of distributions $\mathscr{P}(s, \varpi(s))$ and the constraint on the KL divergence. The entropy uncertainty set $\mathscr{P}_{\mathscr{E}}$ over transition probabilities can be described as follows:

$$\mathscr{P}_{\mathscr{E}} = \Big\{ p_{s,\varpi(s)} \in \mathscr{P}(s, \varpi(s)) : D(p_{s,\varpi(s)} || q_{s,\varpi(s)}) \leq \beta \Big\}.$$

To summarize, the entropy uncertainty set is a probabilistic approach that captures the uncertainty in transition probabilities by considering a set of distributions constrained by the KL divergence. The goal is to identify a policy that performs well under this uncertainty and provides robustness in the face of unknown or varying transition probabilities. This approach allows decision-makers to account for the variability in the system and find policies that perform well in uncertain environments.

4.3.3.1 The Dual Problem

Suppose a state $s \in \mathscr{S}$ and $n = |\mathscr{S}|$ is finite, then the Lagrangian $\mathscr{Q} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ associated with the inner problem can be written as

$$\mathscr{Q}(V, \xi, \nu, \lambda) = p^T V - \xi p + \nu(1 - p^T 1) + \lambda \Big[\sum_{s \in \mathscr{S}} p_{s,\varpi(s)} \log \frac{p_{s,\varpi(s)}}{q_{s,\varpi(s)}} - \beta \Big],$$

where ξ, ν, λ are the Lagrange multipliers. The optimal $p^*_{s,\varpi(s)} = \arg\inf_p \mathscr{Q}(V, \xi, \nu, \lambda)$ is readily be obtained by solving $\partial \mathscr{Q} / \partial p = 0$, which results in

$$p^*_{s,\varpi(s)} = q_{s,\varpi(s)} \exp \Big\{ \frac{\nu - V(s) + \xi(s)}{\lambda} - 1 \Big\}.$$

Plugging the value of $p^*_{s,\varpi(s)}$ back the equation for $\mathscr{Q}(V, \xi, \nu, \lambda)$. By standard duality argument, the inner problem is equivalent to its dual:

$$\max_{\lambda > 0, \nu} \nu - \beta\lambda - \lambda \sum_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp \Big(\frac{\nu - V(s)}{\lambda} - 1 \Big).$$

Setting the derivative with respect to ν to zero, we obtain the optimality condition

$$\sum_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp \Big(\frac{\nu - V(s)}{\lambda} - 1 \Big) = 1$$

from which we derive

$$\nu = \lambda - \lambda \log \Big(\sum_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp\{-\frac{V(s)}{\lambda}\} \Big).$$

The optimal distribution is

$$p^*_{s,\varpi(s)} = \frac{q_{s,\varpi(s)} \exp\{-V(s)/\lambda\}}{\sum\limits_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp\{-V(s)/\lambda\}}.$$

We reduce the inner problem to a one-dimensional optimization problems

$$\max_{\lambda > 0} \sigma(\lambda),$$

where σ is the concave function

$$\sigma(\lambda) = -\lambda \log \Big(\sum_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp\{-\frac{V(s)}{\lambda}\} \Big) - \beta\lambda.$$

4.3.3.2 The Bisection Algorithm

The concave function σ has the following properties :

1 $\forall \lambda \geq 0, V_{\min} - \beta\lambda \leq \sigma(\lambda) \leq q^T V - \beta\lambda$, where $V_{\min} = \min\limits_{s \in \mathscr{S}} V(s)$.

2 $\sigma(\lambda) = V_{\min} - (\beta + \log Q(V))\lambda + o(\lambda)$, where $Q(V) = \sum_{s:V(s)=V_{\min}} q_{s,\varpi(s)} = P(V = V_{\min})$

Hence, $\sigma(0) = V_{\min}$ and $\sigma'(0) = -\beta - \log Q(V)$.
In addition, at infinity the expansion of σ is

$$\sigma(\lambda) = q^T V - \beta\lambda + o(1).$$

If $V(s) = V_{\min}$ for every $s \in \mathscr{S}$, the result holds, with $Q(V) = Q(V_{\min}\mathbf{1}) = 1$. Assume that there exists $s \in \mathscr{S}$ such that $V(s) \geq V_{\min}$. We have

$$\begin{aligned}
\sigma(\lambda) &= -\lambda \log \Big(e^{-\frac{V_{\min}}{\lambda}} \sum_{s \in \mathscr{S}} q_{s,\varpi(s)} \exp\Big(\frac{V_{\min} - V(s)}{\lambda}\Big)\Big) - \beta\lambda \\
&= V_{\min} - \beta\lambda - \lambda \log \Big(\sum_{s:V(s)=V_{\min}} q_{s,\varpi(s)} + \sum_{s:V(s)\geq V_{\min}} q_{s,\varpi(s)} \exp\Big(\frac{V_{\min} - V(s)}{\lambda}\Big)\Big) \\
&= V_{\min} - \beta\lambda - \lambda \log \Big(Q(V) + O(e^{-t/\lambda}) \Big) \\
&= V_{\min} - (\beta + \log Q(V))\lambda - O(\lambda e^{-t/\lambda}),
\end{aligned}$$

where $t = V_s - V_{\min} > 0$ and V_s is the smallest $V(s) > V_{\min}$, $\forall s \in \mathscr{S}$.

The expansion of σ at infinity provides

$$\sigma(\lambda) = -\beta\lambda - \lambda \log\Big(\sum_{s\in\mathscr{S}} q_{s,\varpi(s)}(1 - \frac{V(s)}{\lambda} + o(\lambda))\Big) = q^T V - \beta\lambda - o(1).$$

The bisection algorithm can be started with the lower bound $\lambda_- = 0$. An upper bound can be computed by finding a solution to the equations $\sigma(0) = q^T V - \beta\lambda$, which yields the initial upper bound $\lambda_+ = (q^T V - V_{\min})/\beta$. By concavity, a maximizer exists in the interval $[0, \lambda_+]$.

Algorithm 3 Bisection Algorithm

1 Set $\lambda_+ = (q^T V - V_{\min})/\beta$ and $\lambda_- = 0$. Let $\delta > 0$ be a small convergence parameter.

2 while $\lambda_+ - \lambda_- > \delta(1 + \lambda_- + \lambda_-)$, repeat

a) set $\lambda = (\lambda_+ + \lambda_-)/2$.

b) compute the gradient of σ at λ.

c) if $\sigma'(\lambda) < 0$, set $\lambda_+ = \lambda$; otherwise, $\lambda_- = \lambda$.

d) go to 2 a).

Nilim and El Ghaoui (2005) and Iyengar (2005) have shown that the robust DP problem can be solved by robust finite and infinite horizon DP algorithm and Robust Policy Iteration Algorithm (RPIA) via Bisection algorithm, respectively. Furthermore, Kaufman and Schaefer (2013) propose a robust modified policy iteration algorithm to solve the robust DP problem, which successively approximates the optimal value function and updates the decision rules at each iteration. However, it should be noted that these algorithms may have limitations when applied to real-world problems. For instance, the finite horizon DP algorithm assumes prior knowledge of the terminal value function, which may not be available in practical scenarios. Similarly, the infinite horizon DP algorithm assumes a stationary decision rule, which may not be feasible or practical in real-world applications. Therefore, when applying these algorithms to real-world problems, modifications and adaptations are often necessary to overcome these limitations and make them applicable to specific problem contexts. Additionally, it is important to consider that the robust dynamic programming algorithms can be computationally expensive when state space is large. The calculation of the

optimal value function at each step often leads to overly conservative solutions, which may not be suitable for real-time applications with strict time constraints. In the next section, we will propose a novel approach to address this issue from Reinforcement Learning perspective that strikes a balance between computational efficiency and robustness, allowing for more practical and real-time applicability.

4.4 Risk-sensitive RL with Risk Envelope

Reinforcement Learning (RL) has achieved remarkable success in solving diverse decision-making problems in finance. However, real-world applications often involve environments characterized by uncertainty and risk, which can lead to suboptimal performance or failure of RL algorithms. To tackle this challenge, the field of robust RL with a risk-sensitive objective has emerged as a crucial research area. Its primary aim is to develop algorithms that can effectively handle uncertainties and risks in complex environments.

In this section, we introduce a novel approach from a reinforcement learning (RL) perspective that incorporates a coherent risk measure. Our proposed approach revolves around the concept of a risk envelope, which represents the uncertainty set of transition probabilities. By utilizing the risk envelope, we can effectively pursue robust solutions for risk-sensitive objective functions in the presence of uncertainty surrounding the transition probabilities.

4.4.1 Coherent Risk Measure

Consider a probability space $(\Omega, \mathcal{F}, P_\theta)$, where Ω is the set of outcomes, $\mathcal{F}$ is a σ-algebra over Ω, and $P_\theta \in \mathcal{B}$, where $\mathcal{B} := \{\xi : \int_{\omega\in\Omega} \xi(\omega) = 1, \xi \geq 0\}$, is a probability measure over $\mathcal{F}$ parameter $\theta \in \mathbb{R}^K$. Here, K represents the dimension of the policy parameter θ. Denoted by Z the space of random variables $Z : \Omega \to (-\infty, \infty)$ over the probability space $(\Omega, \mathcal{F}, P_\theta)$. A risk measure is a function $\rho : Z \to \mathbb{R}$ that maps an uncertain outcome to the real line e.g., the expectation $\mathbb{E}[Z]$ or the Conditional Value-at-Risk (CVaR). A risk measure is called coherent, if its satisfies the following conditions for all $Z, W \in Z$:

1. *Convexity*: $\forall\lambda \in [0,1], \rho(\lambda Z + (1-\lambda)W) \leq \lambda\rho(Z) + (1-\lambda)\rho(W)$;

2. *Monotonicity*: if $Z \leq W$, then $\rho(Z) \leq \rho(W)$;

3. *Translation invariance*: $\forall a \in \mathbb{R}, \rho(Z + a) = \rho(Z) + a$;

4. *Positive homogeneity*: if $\lambda \geq 0$, then $\rho(\lambda Z) = \lambda \rho(Z)$.

These conditions intuitively ensures the rationality of risk assessments for a single period. Item 1 ensures that the risk associated with an investment can be reduced through diversification. By spreading investments across multiple assets, the overall risk can be mitigated. Item 2 states that an asset with a higher cost in all possible scenarios is inherently riskier. This reflects the intuitive notion that investments with higher potential losses are riskier. Item 3 refers to as cash invariance, implies the deterministic portion of an investment portfolio does not contribute to its risk. This highlights that holding cash, which has no uncertainty, does not introduce additional risk. Item 4 suggests that doubling the position in an asset also doubles its risk. This relationship emphasizes the proportional nature of risk in investments. These conditions together provide a framework for evaluating risk and making informed investment decisions.

In general, the random variable Z is considered as the cost. In RL setting, the coherent risk measure $\rho(Z)$ is referred to as the risk-adjusted value if a random variable Z is interpreted as the future discount reward. This is true if and only if there exists a convex bounded and closed set $\mathcal{U} \in \mathcal{B}$ such that

$$\rho(Z) = \min_{\xi : \xi P_\theta \in \mathcal{U}(P_\theta)} \mathbb{E}_\xi[Z]. \tag{4.21}$$

It illustrates that any risk-adjusted value is an expectation w.r.t. a worst-case density function ξP_θ, i.e., a reweighting of P_θ by ξ, chosen adversarially from a suitable set of test density function $\mathcal{U}(P_\theta)$, referred to as risk envelope. In addition, a risk-adjusted value is uniquely represented by its risk envelope (Tamar et al., 2016). In this study, we assume that the risk envelope $\mathcal{U}(P_\theta)$ is given in a canonical convex programming formulation and satisfies the following conditions.

Assumption 4.4.1 *For any given policy parameter $\theta \in \mathbb{R}^K$, the risk envelope $\mathcal{U}$ of a coherent*

risk measure can be written as

$$\mathcal{U}(P_\theta) = \Big\{\xi P_\theta : g_e(\xi, P_\theta) = 0, \forall e \in \mathcal{E}, f_\iota(\xi, P_\theta) \leq 0, \forall \iota \in \mathcal{I}, \sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) = 1, \xi(\omega) \geq 0\Big\}, \tag{4.22}$$

where each constraint $g_e(\xi, P_\theta)$ is an affine function in ξ, each constraint $f_\iota(\xi, P_\theta)$ is a convex function in ξ, and there exists a strictly feasible point $\bar{\xi}$. $\mathcal{E}$ and $\mathcal{I}$ here denote the finite sets of equality and inequality constraints, respectively. In addition, $f_\iota(\xi, p)$ and $g_e(\xi, p)$ are twice differentiable in p, and there exists $M > 0$ such that

$$\max\left\{\max_{\iota\in\mathcal{I}}\left|\frac{df_\iota(\xi,p)}{dp(\omega)}\right|, \max_{e\in\mathcal{E}}\left|\frac{dg_e(\xi,p)}{dp(\omega)}\right|\right\} \leq M, \forall\omega \in \Omega.$$

From the above assumption, it implies that the risk envelope $\mathcal{U}(P_\theta)$ is known in an explicit form. Note that in the case of a finite probability space, ρ is a coherent risk measure if and only if $\mathcal{U}(P_\theta)$ is a convex and compact set.

4.4.2 Dynamic Risk Measures

Dynamic risk measures are tools utilized in finance and economics to evaluate risk over a sequence of time points. Unlike static risk measures which consider risk at a single time point, dynamic risk measures account for the time-varying nature of risk, especially pertinent in investment decisions and risk management scenarios. In static risk measures, we might consider the risk of an investment at a specific point in time. However, in reality, investments often span over periods of time and the risk associated with them may vary over this duration. In addition, dynamic risk measures introduce the concept of time-consistent. This implies that if a strategy is deemed risk-optimal at the onset of a period, it should remain considered as such at any subsequent point within this period. This property is crucial for ensuring the stability and reliability of investment strategies.

Markov coherent risk measures are a specialized form of dynamic risk measures. It is designed to assess risk in situations where the stochastic processes involved exhibit the Markov property. The Markov property suggests that the future state of a process depends only on its current state and not on its past states. This is a key feature in many economic and financial

systems, making Markov coherent risk measures especially relevant for these applications. In general, Markov coherent risk measures provide a time-consistent, dynamic evaluation of risk. In RL settings, the Markov coherent risk measure $\rho_T(\mathcal{M})$ for a T-period horizon is defined as follows:

$$\rho_T(\mathcal{M}) = r(s_0, a_0) + \gamma\rho\bigg(r(s_1, a_1) + ... + \gamma\rho\Big(r(s_{T-1}, a_{T-1}) + \gamma\rho(r(s_T))\Big)\bigg),$$

where $\{s_0, a_0, i_0, ..., s_{T-1}, a_{T-1}, i_{T-1}, s_T\}$ is a trajectory drawn from an MDP $\mathcal{M}$, γ is the discount factor, and ρ is a static coherent risk measure that satisfies Assumption 4.4.1. Each static coherent risk measure ρ at state $s \in \mathscr{S}$ is induced by the transition probability $P(s'|s) = \sum_{i\in\mathscr{I}} P(s'|s, i)\varpi_\vartheta(i|s)$.

4.4.3 RL with Risk Envelope

We consider an MDP $\mathcal{M}$ with a hidden regime rules ϖ parametrized by ϑ, a policy of action ν parameterized by θ, and a transition probability P, Z may correspond to the cumulative discounted future rewards

$$Z = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t),$$

where $r(s_t, a_t)$ is a bonded and deterministic reward function, and γ is the discount factor. The actual action is chosen by a stochastic policy that depends on hidden regime i_t and current state s_t, that is, $\nu_\theta(a_t|s_t, i_t)$. The hidden regime is estimated based on stochastic hidden regime rules, that is, $i_t = \varpi_\vartheta(i|s_t)$. We also assume that the transition probability in MDP $\mathcal{M}$ depends on the current state and hidden regime mechanisms such as $P(s_{t+1}|s_t, i_t)$. Now, let $T \to \infty$, Markov coherent dynamic risk measure can be defined as follows:

$$\rho_\infty(\mathcal{M}) = \lim_{T\to\infty} \rho_T(\mathcal{M}) = r(s_0, a_0) + \gamma\rho\bigg(r(s_1, a_1) + \gamma\rho\Big(r(s_2, a_2) + ...\Big)\bigg),$$

where ρ is a risk-adjusted value that can be uniquely represented by its risk envelope (4.21). The objective function of the risk-sensitive RL problems can be formulated as

$$\max_\theta J_\theta(s_0) = \max_\theta \rho_\infty(\mathcal{M}).$$

Inspired by Ruszczyński (2010), we define the value function for a fixed $s = s_0$ as follows:

$$V(s) = \rho_\infty(\mathcal{M}) = r(s_0, a_0) + \gamma\rho\Big(r(s_1, a_1) + \gamma\rho(s_2, a_2) + ...\Big). \tag{4.23}$$

Let $\mathcal{B}(\mathscr{S})$ denotes the space of real-valued bounded functions on state space $\mathscr{S}$. The risk-sensitive Bellman operator $T_{\theta,\vartheta}[V] : \mathcal{B}(\mathscr{S}) \to \mathcal{B}(\mathscr{S})$ is defined as

$$T_{\theta,\vartheta}V(s) = \min_{\xi \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma V(\hat{s})\Big|s\Big], \tag{4.24}$$

where $\hat{s} \in \mathscr{S}$ and $\hat{a} \in \mathscr{A}$ are random variables such that $(\hat{a}, \hat{s}) \sim P_{\theta,\vartheta}(\hat{a}, \hat{s}|s) = \sum_{i\in\mathscr{I}} \varpi_\vartheta(i|s)$ $\nu_\theta(\hat{a}|s, i)P(\hat{s}|s, i)$, and $i \in \mathscr{I}$ denotes the estimated hidden regime in the system.

Proposition 4.4.1 *The risk-sensitive Bellman operator is a γ-contraction w.r.t ∞-norm.*

Proof. Fix $s \in \mathscr{S}$ and $i \in \mathscr{I}$, we assume that $T_{\theta,\vartheta}Y(s) \geq T_{\theta,\vartheta}Z(s)$. Choose some $\epsilon \geq 0$ and $\xi P_{\theta,\vartheta}(\cdot,\cdot|s) \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot.\cdot\,|s))$ such that

$$\mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Z(\hat{s})\Big|s\Big] \leq \min_{\xi \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Z(\hat{s})\Big|s\Big] + \epsilon$$

By definition, we have

$$\min_{\xi \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Y(\hat{s})\Big|s\Big] \leq \mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Y(\hat{s})\Big|s\Big].$$

Then, we have

$$\begin{aligned} 0 \leq T_{\theta,\vartheta}Y(s) - T_{\theta,\vartheta}Z(s) &\leq \mathbb{E}_{\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Y(\hat{s})\Big|s\Big] - \Big(\mathbb{E}_{\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a}) + \gamma Z(\hat{s})\Big|s\Big] - \epsilon\Big) \\ &\leq \gamma\Big|\Big|Y(s) - Z(s)\Big|\Big|_\infty \end{aligned}$$

where $\hat{P}_{\theta,\vartheta}(\cdot,\cdot|s)$ is the transition matrix. Conversely, if $T_{\theta,\vartheta}Y(s) \leq T_{\theta,\vartheta}Z(s)$, following the same procedure, we can obtain

$$0 \leq T_{\theta,\vartheta}Z(s) - T_{\theta,\vartheta}Y(s) \leq \gamma\Big|\Big|Y(\hat{s}) - Z(\hat{s})\Big|\Big|_\infty.$$

Thus, we can conclude that

$$\Big|\Big|T_{\theta,\vartheta}Y - T_{\theta,\vartheta}Z\Big|\Big|_\infty \leq \gamma\Big|\Big|Y - Z\Big|\Big|_\infty.$$

This establishes the risk-sensitive bellman operator is a γ-contraction w.r.t. ∞-norm. This completes the proof. ■

Proposition 4.4.1 demonstrate that the risk-sensitive bellman equation is a γ-contraction operator on a Banach space. Therefore, by the Banach fixed point theorem, the operator equation $T_{\theta,\vartheta}V = V$ has a unique fixed point. According to Theorem 4 in Ruszczyński (2010), the fixed point is equal to the value function defined in (4.23), i.e. $V(s) = \rho_\infty(\mathcal{M})$. This value function assigns to each state a particular value that encodes the long-term risk of the system staring from that state. However, when the state space $\mathscr{S}$ is large, exact enumeration of the Bellman equation is intractable due to calculating $V(s)$ for every state $s \in \mathscr{S}$ is prohibitively computationally expensive, and a lower dimensional approximation of V is sought (Tamar, Mannor, & Xu, 2014).

4.4.4 Gradient of Value Function

According to the Assumption 4.4.1, $\rho(Z)$ is a risk-adjusted value that can be calculated by equation (4.21). The Lagrangian function of (4.21), denoted by $\mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$, can be written as

$$\begin{aligned}\mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = &\sum_{\omega\in\Omega} \xi(\omega) P_{\theta,\vartheta}(\omega) Z(\omega) + \lambda^{\mathcal{P}}\Big(\sum_{\omega\in\Omega} \xi(\omega) P_{\theta,\vartheta}(\omega) - 1\Big) \\ &+ \sum_{e\in\mathcal{E}} \lambda^{\mathcal{E}}(e) g_e(\xi, P_{\theta,\vartheta}) + \sum_{\iota\in\mathcal{I}} \lambda^{\mathcal{I}}(\iota) f_\iota(\xi, P_{\theta,\vartheta}).\end{aligned}$$

It also illustrates that $\mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is concave in ξ and convex in $(\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$. Thus, it implies that the strong duality holds, that is,

$$J_{\theta,\vartheta}(Z) = \max_{\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}\geq 0} \min_{\xi\geq 0} \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \min_{\xi\geq 0} \max_{\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}\geq 0} \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}).$$

Assumption 4.4.1 also depicts that the $\mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is Lipschitz and an absolutely continuous function in θ, and there exists a non-empty set of saddle points $\mathcal{S}$. Thus, $\nabla_\theta \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})$ is continuous and bounded. For every selection of saddle points $(\xi^*, \lambda^{*,\mathcal{P}}, \lambda^{*,\mathcal{E}}, \lambda^{*,\mathcal{I}}) \in \mathcal{S}$, using the envelope theorem from Milgrom and Segal (2002) for saddle-point problems, we have

$$\nabla_\theta \min_{\xi\geq 0} \max_{\lambda^{\mathcal{P}},\lambda^{\mathcal{E}},\lambda^{\mathcal{I}}\geq 0} \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}) = \nabla_\theta \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}})\Big|_{\xi^*,\lambda^{*,\mathcal{P}},\lambda^{*,\mathcal{E}},\lambda^{*,\mathcal{I}}}.$$

Let $(\xi^*, \lambda^{*,\mathcal{P}}, \lambda^{*,\mathcal{E}}, \lambda^{*,\mathcal{I}})$ be the saddle point for the state $s \in \mathscr{S}$. In many common coherent risk measures such as CVaR and semi-deviation, they provide closed-form formulas for ξ^* and KKT multipliers $(\lambda^{*,\mathcal{P}}, \lambda^{*,\mathcal{E}}, \lambda^{*,\mathcal{I}})$. Before analyzing the gradient of value function, we have the following standard assumption.

Assumption 4.4.2 *The likelihood ratio $\nabla_\theta \log \nu_\theta(a|s,i)$ is well-defined and bounded for all $s \in \mathscr{S}$, $a \in \mathscr{A}$ and $i \in \mathscr{I}$.*

As previously discussed, the Bellman operator $T_{\theta,\vartheta}V = V$ has a unique fixed point, and the fixed point corresponds exactly to value function defined in (4.23). Consequently, we can employ the risk-sensitive Bellman equation

$$V(s) = \min_{\xi \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim \xi P_{\theta,\vartheta}(\cdot,\cdot|s)} \Big[r(s,\hat{a}) + \gamma V(\hat{s}) \Big| s \Big].$$

to derive the gradient of value function $\nabla_\theta V(s)$.

Theorem 4.4.3 *Under Assumptions 4.4.1 and 4.4.2, the gradient of value function w.r.t θ can be deduced as follows.*

$$\nabla_\theta V(s) = \mathbb{E}_{\xi^* P_{\theta,\vartheta}(\cdot,\cdot|s)} \Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \nu_\theta(a_t|s_t,i_t) h(s_t,a_t,i_t) \Big| s_0 = s \Big], \tag{4.25}$$

where $\mathbb{E}_{\xi^* P_{\theta,\vartheta}}[\cdot]$ denotes the expectation under worst-case scenario w.r.t. trajectories generated by a MDP with hidden regime rule $i \sim \varpi_\vartheta(\cdot|s)$, action probability $a \sim \nu_\theta(\cdot|s,i)$, and transition probability $P(\cdot|s,i)\xi^*(i,\cdot)$. The stage-wise function h is given by

$$\begin{aligned} h(s,a,i) =& r(s,a) + \sum_{s' \in \mathscr{S}} P(s'|s,i)\xi^*(i,s') \Big[\gamma V(s') + \lambda^{*,\mathcal{P}} + \sum_{\iota \in \mathcal{I}} \lambda^{*,\mathcal{I}}(\iota) \frac{df_\iota(\xi^*, P_{\theta,\vartheta})}{dp(a,s')} \\ &+ \sum_{e \in \mathcal{E}} \lambda^{*,\mathcal{E}}(e) \frac{dg_e(\xi^*, P_{\theta,\vartheta})}{dp(a,s')} \Big]. \end{aligned}$$

Proof. Similar to the proof of Theorem V.4 in Tamar et al. (2016), by the strong duality result, we have

$$\min_{\xi \in \mathcal{U}(s, P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim \xi P_{\theta,\vartheta}(\cdot,\cdot|s)} \Big[r(s,\hat{a}) + \gamma V(\hat{s}) \Big| s \Big] = \min_{\xi \geq 0} \max_{\lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}} \geq 0} \mathcal{L}_{\theta,\vartheta}(\xi, \lambda^{\mathcal{P}}, \lambda^{\mathcal{E}}, \lambda^{\mathcal{I}}).$$

Therefore, the gradient formula can be written as

$$\begin{aligned}\nabla_\theta V(s) &= \nabla_\theta \min_{\xi\in\mathcal{U}(s,P_{\theta,\vartheta}(\cdot,\cdot|s))} \mathbb{E}_{(\hat{a},\hat{s})\sim\xi P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[r(s,\hat{a})+\gamma V(\hat{s})\Big|s\Big] \\ &= \sum_{i\in\mathscr{I},a\in\mathscr{A},s'\in\mathscr{S}} \varpi_\vartheta(i|s)\nu_\theta(a|s,i)\Big\{P(s'|s,i)\xi^*(i,s')\nabla_\theta\gamma V(s')+\nabla_\theta\log\nu_\theta(a|s,i)h(s,a,i)\Big\},\end{aligned}$$

where

$$\begin{aligned}h(s,a,i) =& r(s,a)+\sum_{s'\in\mathscr{S}}P(s'|s,i)\xi^*(i,s')\Big[\gamma V(s')+\lambda^{*,\mathcal{P}}+\sum_{\iota\in\mathcal{I}}\lambda^{*,\mathcal{I}}(\iota)\frac{df_\iota(\xi^*,P_{\theta,\vartheta})}{dp(a,s')} \\ &+\sum_{e\in\mathcal{E}}\lambda^{*,\mathcal{E}}(e)\frac{dg_e(\xi^*,P_{\theta,\vartheta})}{dp(a,s')}\Big].\end{aligned}$$

By defining $\hat{h}(s,a,i)=\nabla_\theta\log\nu_\theta(a|s,i)h(s,a,i)$ and unfolding the recursion, the above expression implies

$$\begin{aligned}\nabla_\theta V(s_0) =& \sum_{a_0\in\mathcal{A},i_0\in\mathscr{I}}\hat{h}(s_0,a_0,i_0)+\gamma\sum_{s_1\in\mathscr{S}}P(s_1|s_0,i_0)\xi(i_0,s_1)\Big[\sum_{a_1\in\mathcal{A},i_1\in\mathscr{I}}h(s_1,i_1,a_1) \\ &+\gamma\sum_{s_2\in\mathscr{S}}P(s_2|s_1,i_1)\xi(i_1,s_2)\nabla_\theta V(s_2)\Big]\end{aligned}$$

Now since $\nabla_\theta V(s)$ is continuously differentiable with bounded derivatives, when $t\to\infty$, one obtains $\gamma^t\nabla_\theta V\to 0$ for any $s\in\mathscr{S}$. Therefore, by bounded convergence theorem, $\lim_{t\to\infty}\rho(\gamma^t\nabla_\theta V(s))=0$, which implies the above expression can be estimated as

$$\nabla_\theta V(s)=\mathbb{E}_{\xi^*P_{\theta,\vartheta}(\cdot,\cdot|s)}\Big[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta\log\nu_\theta(a_t|s_t,i_t)h(s_t,a_t,i_t)\Big|s_0=s\Big].$$

This completes the proof. ■

It is worth noting that the policy gradient of the Markov-coherent dynamic risk measure $\rho_\infty(\mathcal{M})$, i.e. $\nabla\rho_\infty(\mathcal{M})$ is equivalent to the risk-neutral value function of policy θ in a MDP with a stage-wise function $\nabla_\theta\log\nu_\theta(a|s,i)h_\theta(s,a,i)$ that is both well-defined and bounded. The MDP also has an action probability $\nu_\theta(\cdot|s,i)$ and a transition probability $P(\cdot|s,\cdot)\xi(i,\cdot)$. Therefore, if the saddle points are known and the state space is not excessively large, we can calculate the gradient by using a policy evaluation algorithm. However, when the state space is too large, the exact computation of ∇V_θ by policy evaluation becomes impractical.

4.5 Implementation

Handling the calculation of the gradient for the coherent risk measure becomes challenging when the sample space is large. In such cases, the most popular approach is to use a natural Monte-Carlo (MC) estimation algorithm. The MC estimation algorithm involves generating a set of sample paths by running the MDP under the current policy. The gradient is then estimated using a sample average of the instantaneous gradients along the generated sample paths. In this section, we present a novel reinforcement learning (RL) algorithm aimed at constructing a risk-sensitive policy under the worst-case scenario.

We are examining a portfolio management problem in which the agent is mandated to choose a portfolio allocation strategy that maximizes the expected alpha-percentile of the future discounted return. In this problem, we adopt a hidden-regime dependent MDP to formulate this problem, which is described in section 4.3. In particular, we hypothesize that the transition probability in MDP is contingent on the current state and the hidden regime of the financial market, represented as $P(s_{t+1}|s_t, i_t)$. The future discounted return is defined as the discounted sum of the rewards, that is, $R = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$, where $r(s_t, a_t)$ is the reward at time t and $\gamma \in (0, 1)$ is the discount factor. We adopt the expected alpha-percentile as the risk-sensitive objective with confidence level $\alpha \in (0, 1]$, which is defined as follows:

$$J_\alpha(R) = \mathbb{E}[R|R \leq \mathit{percentile}(\alpha)].$$

The decreasing nature of J_α with respect to α is a well-established fact. This can be interpreted as the worst-case expected value of R given the α-portion of the left tail distribution. Furthermore, it is important to note the expected alpha-percentile objective function satisfies the necessary conditions for a coherent risk measure. This property is supported by the alternative dual representation of a coherent risk measure, as stated by Artzner, Delbaen, Eber, and Heath (1999):

$$J_\alpha(R) = \min_{\xi: \xi P_{\theta,\vartheta} \in \mathcal{U}(P_{\theta,\vartheta})} \mathbb{E}_{\xi P_{\theta,\vartheta}}\Big[R\Big], \tag{4.26}$$

where $\mathbb{E}_{\xi P_{\theta,\vartheta}}[R]$ denotes the ξ-weighted expectation of R, and the risk envelope $\mathcal{U}$ is defined as:

$$\mathcal{U}(P_{\theta,\vartheta}) = \Big\{\xi P_{\theta,\vartheta} : \xi(i, s') \in [0, \alpha^{-1}], \sum_{s',i} \xi(i, s') P_{\theta,\vartheta}(s'|s, i) = 1\Big\}. \tag{4.27}$$

In this context, ξ is selected from the risk envelope to adjust the worst-case probability density, reflecting the worst-case scenario that can arise under a perturbed distribution $\xi P_{\theta,\vartheta}$. Our goal is to maximize the performance in the worst case scenarios, which can be achieved by solving the following optimization problem

$$\max_{\theta} \min_{\xi \in \mathcal{U}(P_{\theta,\vartheta})} \mathbb{E}_{\xi P_{\theta,\vartheta}}\Big[R\Big]. \tag{4.28}$$

The cornerstone of our algorithm involves transforming the continuous state of stock closing prices into a finite state space. This transformation is facilitated through Equal-Frequency Discretization, a technique that segregates the continuous price values into a distinct number of states, each containing a roughly equal range of observed prices. Consider a series of historical closing prices symbolized as $\{v_0^c, v_1^c, v_2^c, ..., v_m^c\}$. We initiate the process by calculating relative price change, that is, taking a ratio of the current closing price to the previous closing price, e.g., $y_t = v_t^c / v_{t-1}^c$ for $t = 1, 2, 3, ..., m$. Next, we sort the relative price changes in ascending order, denoted as $\{y_{(1)}, y_{(2)}, ..., y_{(m)}\}$. We then divide the sorted price changes into N equal-sized intervals, denoted as $\{I_1, I_2, ..., I_N\}$. Assuming each state encompasses w price changes, the price change at each time is accordingly assigned to the interval to which it belongs. For example, if we treat a state as a sequence of price changes spanning five days, every price change within this 5-day duration is allocated to its corresponding interval. As a result, we acquire a finite state space containing 5^N unique states. Through Equal-Frequency Discretization, we are successful in transforming the continuous stream of stock closing prices into a finite collection of discrete states. This conversion greatly facilitates the subsequent analytical steps while also enabling a more sophisticated decision-making process.

The starting point of our algorithm is to estimate the empirical transition probabilities using the equation: $P(s'|s) = \sum_i \varpi_\vartheta(i|s)\hat{P}(s'|s,i)$. Here, $\hat{P}(s'|s,i)$ represents the estimated probability of transitioning from state s to next state s' given the hidden regime i, and $\varpi_\vartheta(i|s)$ is a network that predicts the hidden regime based on the current state s_t. Specifically, $\varpi_\vartheta(i|s)$ outputs the probability of the hidden regimes i_t in state s_t. The empirical transition probability $\hat{P}(s'|s,i)$ is estimated by the following equation:

$$\hat{P}(s'|s,i) = \frac{N(s,i,s')}{N(s,i)},$$

where $N(s, i, s')$ is the number of times the states has transitioned from state s to state s' under the hidden regime i, and $N(s, i)$ is the total occurrence of state s under the hidden regime i.

In the process of training our hidden regime rule, denoted as ϖ_ϑ, we incorporate the use of a Hidden Markov Model (HMM). This model serves a crucial role in furnishing a label that assists in training our network. The HMM is an excellent fit for our scenario due to its proficient handling of situations involving hidden or latent variables. This distinctive capability makes the HMM a robust instrument for labeling our states, which subsequently feeds into the training of our network ϖ_ϑ. Once the HMM is trained, it generates a sequence of hidden states or labels corresponding to each observed state. These labels are then used to train our network ϖ_ϑ, effectively enabling the network to learn and emulate the underlying regime rules encoded by the HMM. Through this procedure, the network ϖ_ϑ becomes adept at identifying the hidden regimes and making corresponding decisions based on the identified regime, thereby enhancing the overall performance of our proposed framework.

To tackle the inner problem, we need to find the optimal value function $V(s')$ for all possible states s' in finite state space. This value function approximates the expected cumulative rewards starting from state s' and following the current policy. To mitigate uncertainty in transition matrix rows, we assume uncertainty set U for all the rows in the transition matrix and apply the robust Bellman recursion equation to decipher the inner problem. The robust Bellman recursion equation is articulated as:

$$V(s) = \min_{\xi \in \mathcal{U}(P_{\theta,\vartheta})} \mathbb{E}_{\xi P_{\theta,\vartheta}} \Big[r_t + \gamma V(s') \Big],$$

where the value function $V(s)$ is defined as

$$V(s) = \mathbb{E}[R | R \leq \mathit{percentile}(\alpha), s].$$

In this context, ξ represents a specific set of transition probabilities encapsulated within the uncertainty set $\mathcal{U}$, and $\mathbb{E}_{\xi P_{\theta,\vartheta}}$ denotes the expectation associated with the distribution defined by $\xi P_{\theta,\vartheta}$. We take the minimum over all candidate value functions to ensure that the value function is robust to the uncertainty in the transition matrix. The uncertainty set has been defined in (4.26) and (4.27). At each timestep, the inner problem is solved by using

the CVXPY optimization model in python with the MOSEK solver. Within our proposed framework, the target value y_t is computed as:

$$y_t = \inf_{\xi \in \mathcal{U}(P_{\theta,\vartheta})} \mathbb{E}_{\xi P_{\theta,\vartheta}} \Big[r_t + \gamma V(s') \Big], \tag{4.29}$$

where ξ can be solved by the MOSEK solver at each timestep, and the value function $V(s')$ can be acquired by utilizing the Critic network with the equation $V(s') = \sum_{a,i} \varpi_\vartheta(i|s')\nu_\theta(a|s',i)Q(s',a)$. The Critic network $Q(s,a)$ is refined by minimizing the loss function:

$$L = \frac{1}{N} \sum_{t=1}^{N} (y_t - Q(s_t, a_t))^2.$$

Subsequently, the actor network ν_θ can be updated using Theorem 4.4.3. The completed robust actor-critic with hidden regime rules algorithm is outlined in 4.

To commence the process, parameters θ and ω of the Actor and Critic networks are initialized. Each episode commences with initializing the state s and the replay buffer D. Subsequently, the execution phase then launches with the transformation of the current state into a finite state, represented by s_t. The hidden regime i_t is estimated by using the hidden regime rules ϖ_ϑ. A vital step is the estimation of the empirical transition matrix $\hat{P}(s'|s,i)$, which is computed using the pretrained hidden regime rules ϖ_ϑ. An action a_t is sampled based on the current state s_t and hidden regime i_t, then executes it. The estimated hidden regime i_t leads to the generation of a new state s_{t+1} and an associated reward r_t from environment. All these elements, i.e., the state, action, and reward, are stored in the experience replay buffer D. The following stage entails extracting a sequence of N transitions from the replay buffer. For each transition (s_t, a_t, r_t), a target value y_t is computed. This calculation is facilitated by solving an inner problem using a designated equation for the target value. Pivotal to the learning process are the updates to the Critic and the Actor. The Critic is fine-tuned by minimizing the loss function L, which encapsulates the variance between the target value y_t and the value derived from the Critic network $Q_\omega(s_t, a_t)$. The Actor, contrarily, is refreshed using the policy gradient outlined in Theorem 4.4.3. The algorithm cyclically processes these steps until the end of the episode. As the algorithm advances, the Critic and Actor networks incrementally evolve, yielding increasingly optimal solutions within the given environment. This constant refinement is a testament to the prowess of reinforcement learning, where

learning from experiences fuels enhanced decision-making. The workflow of the proposed algorithm is illustrated in Figure 4.1.

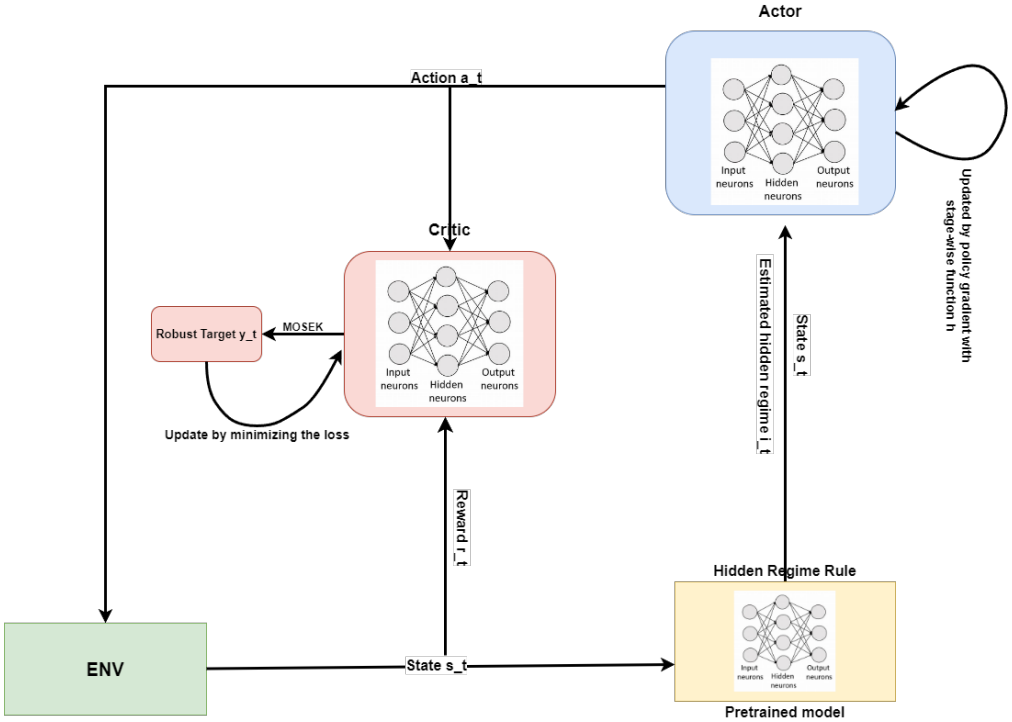


Figure 4.1: Workflow of the proposed algorithm

4.6 The Experiment Results

In this section, we verify the effectiveness of our proposed method by using real-world data and simulated data. Our findings underscore the critical role of the hidden regime in the financial market. By accounting for this, market fluctuations can be more precisely discerned, thereby facilitating more robust decision-making. Furthermore, we demonstrate our approach can improve policy robustness, enabling it to be effectively adjusted to various market environments.

Algorithm 4 Actor-Critic Model with hidden regime rules

Require: Environment, Actor network ν_θ, Critic network Q_ω, pretrained hidden regime rules ϖ_ϑ, and discount factor γ.

Initialize parameters θ and ω of Actor and Critic networks

for each episode **do**

  Initialize state s

  Initialize replay buffer D

  Convert continuous state to finite state space

  Estimated empirical transition matrix $\hat{P}(s_{t+1}|s_t) = \sum_{i_t} \varpi_\vartheta(i_t|s_t)\hat{P}(s_{t+1}|s_t, i_t)$

  while episode not over **do**

    Sample an action $a_t \sim \sum_{i_t} \varpi_\vartheta(i_t|s_t)\nu_\theta(a_t|s_t, i_t)$ based on current state s_t, execute a_t, and get new state s_{t+1} and reward r_t

    Store (s_t, a_t, r_t) in experience replay buffer D.

    Sample a sequence of N transitions $(s_{1:N}, a_{1:N}, r_{1:N})$ from replay buffer.

    Compute the target value y_t for each transition (s_t, a_t, r_t) by solving the inner problem using equation (4.29).

    Update the critic by minimizing the loss: $L = \frac{1}{N}\sum_t (y_t - Q_\omega(s_t, a_t))^2$.

    Update the actor by using the policy gradient in theorem 4.4.3:

$$\nabla_\theta J = -\frac{1}{N}\sum_{t=1}^{N} \gamma^t \nabla_\theta \log \nu_\theta(a_t|s_t, i_t) h(s_t, a_t, i_t), \tag{4.30}$$

    where $h(s_t, a_t, i_t) = r(s_t, a_t) + \sum_{s' \in \mathscr{S}} P(s'|s_t, i_t)\xi^*(i_t, s')\Big[\gamma V(s') + \lambda^{*,\mathcal{P}}\Big]$.

  end while

end for

4.6.1 Data

Real-world Data

We utilized real-world data in the form of Apple Inc. stock prices from the U.S. market, sourced from Yahoo Finance. This dataset spans a decade and encompasses opening, highest, lowest, and closing prices (OHLC). Given Apple's standing as one of the world's most lucrative and influential companies, its stock prices, which can significantly influence the overall stock market, are closely monitored by investors and analysts. Hence, this dataset is frequently employed in financial research and analysis. By analyzing this dataset, we can gain deeper insights into the role of covert regimes in decision-making.

Simulated Data

In addition to real-world data, we used simulated data to explore the variation in stock prices under differing levels of volatility. We produced a decade's worth of data, including the opening, highest, lowest, and closing prices. A Geometric Brownian Motion (GBM) model was deployed to simulate the stock prices. Due to its capacity to reflect the unpredictability and volatility of stock prices, the GBM model is commonly applied in financial modeling. In our simulation, we modified the parameters of the GBM model to generate stock prices exhibiting varying degrees of volatility. Specifically, we generated two sets of simulated data with annual volatilities of 1.5% and 15%, representing low and high volatility scenarios, respectively. These simulated datasets were then employed to validate our analytical models and assess their performance under diverse market conditions.

4.6.2 Problem Formulation

For this experiment, we consider the individual asset that consists of the opening, highest, lowest, and closing prices at each time. The closing price is particularly critical in the stock market. As the final price at which a stock is traded during regular trading hours, it's widely utilized in calculating various stock market indicators such as the price-to-earnings ratio and the dividend yield. Given its extensive application in stock market analysis, we've selected

the closing price to represent asset prices in our study.

State

For the t-th trading period, the agent observes the history of the closing prices from $v^c_{t-(m+1)}$ to v^c_t, that is $\{v^c_{t-(m+1)}, v^c_{t-m}, ..., v^c_t\}$, and m represents the window size. In each trading period, these observations need to be transformed into a finite state, $s_t \in \mathscr{S}$. In this context, $\mathscr{S}$ denotes the finite state space. To characterize the state s_t at the t-th trading period, we use the ratio of the current closing price to the previous closing price. This is expressed as $\{\frac{v^c_{t-m}}{v^c_{t-(m+1)}}, ..., \frac{v^c_t}{v^c_{t-1}}\}$. To discretize these continuous variables, we employ the Equal-Frequency Discretization method, which we described in section 4.5. This method divides the range of a continuous variable into N intervals, each containing an equal number of observations. The boundaries of these intervals are dictated by the quantiles of the variable in question, while the user determines the number of intervals, N. At each time t, the finite state s_t is a vector of length m and each element of the vector is the least upper bound of the corresponding interval. In our experiment, we set both N and m to 5.

Action and Reward

At this juncture, the agent can select action a_t from the finite action space, denoted as $\mathcal{A}$. To simplify the computation, we define the finite action space as $\mathcal{A} = \{0, 0.5, 1\}$, symbolizing actions of holding all cash, maintaining half of the assets and half cash, and holding all assets, respectively. Upon selecting an action a_t, the agent receives a reward r_t. We assume p_t is the portfolio value at time t. Therefore, the reward function r_t can be defined as follows: In this experiment, the reward function r_t accounts for the transaction cost and is defined as follows:

$$r_t = \log \frac{p_t}{p_{t-1}} = \log((1 - C_t)\frac{v^c_t}{v^c_{t-1}} \cdot a_{t-1}),$$

where C_t is the transaction cost in t-th trading period, and a_{t-1} is portfolio allocation at beginning of the t-th trading period. The transaction cost C_t is defined as follows:

$$C_t = c\Big|a_t - a_{t-1}\Big|,$$

where c is the transaction cost rate. For a more comprehensive understanding of the problem settings, please refer to M. Wang and Ku (2022).

4.6.3 The Results

Our approach was assessed through a series of experiments which explored models assuming no hidden regimes, as well as two and three hidden regimes in the financial market. For the real-world data, a Hidden Markov Model (HMM) was utilized to discern the hidden regimes in the financial market. Subsequently, a neural network was used to estimate the probability of these hidden regimes based on a series of closing prices. To identify the regimes in the financial market, we used an equivalent period of historical S&P 500 SPY ETF OHLC data, recognized broadly as an indicator of the United States' overall economic health and investor confidence in the stock market. Applying the HMM model to this dataset allowed us to evaluate the hidden regimes in the market and use the resultant labels as ground truth for training our neural network classifier. For each simulated dataset, we leveraged a sequence of closing prices from the GBM model to evaluate the hidden regimes in the financial market. This enabled us to assess the robustness and effectiveness of our approach under different levels of market volatility. By appraising our approach on both simulated and real-world datasets, we can assess its effectiveness in accurately predicting hidden regimes and making more informed investment decisions based on anticipated market trends. Due to computational constraints, in our experiment, we limited our consideration to an action space consisting of only three actions: holding all, holding half, and not holding any. This restriction allowed us to focus on the core components of our approach while managing computational resources effectively. Despite this limitation, our experiments demonstrate that our approach is adaptable and can be used with different action spaces, providing flexibility for different investment strategies.

4.6.3.1 The Experimental Results for Real-world Data

Figure 4.2 illustrates the worst-case performance ($\alpha = 0.05$) of real-world data Apple Inc. with 5-day window size. In examining the real-world dataset (AAPL), we recommend

exploring models with different numbers of hidden regimes and evaluating their performance under the worst-case scenarios. As shown in Figure 4.2, we have found that no single model consistently outperforms the others. In the early period, the model with two hidden regimes demonstrated superior performance in the worst-case scenario. However, in the later period, the model with three hidden regimes outshined the rest under the worst-case scenario. Additionally, we observed that the model with three hidden regimes yielded the highest Sharpe ratio, while the model without hidden regimes produced the second-highest Sharpe ratio. In general, models that integrate hidden regimes can significantly enhance the performance of robustness, and also can be efficient to formulate for the worst-case scenario. Nevertheless, the optimal number of hidden regimes may fluctuate based on the dataset and the problem at hand. For more details on the performance for real-world data Apple, please refer to Figures C.1- C.3 in the Appendix.

Figure 4.2: The performance of AAPL with different hidden regimes

4.6.3.2 The Experimental Results for Simulated Data

We generate two datasets using the Geometric Bowinan Motion (GBM) model, which have the same expected value but exhibit different volatility. The first dataset is referred to as BCKCXA, the second as BCKCX. BCKCXA has an annual volatility of 1.5%, while BCKCX has a volatility of 15%. Assuming that the financial market possesses two or three hidden regimes, we are able to detect these hidden regimes by applying the HMM model. The closing prices along with hidden regimes of BCKCXA and BCKCX are portrayed in Figures C.4-C.9 in the Appendix C.

Figure 4.3 and Figure 4.4 show the performance of the worst-case ($\alpha = 5\%$) for BCKCXA and BCKCX respectively, each with a 5-day window size. Figure 4.3 demonstrates that in a dataset with lower volatility, the model with three hidden regimes outperforms both the model with two hidden regimes and the model without any hidden regime. Furthermore, the model with two hidden regimes also performs better than the model without any hidden regime. Figure 4.4 demonstrates that, in a dataset with high volatility, the model with three hidden regimes outperforms both the model with two hidden regimes and the model without any hidden regime. Moreover, the model with two hidden regimes also outperforms the model without any hidden regime.

Table 4.2 presents outlines the performance of different models that are tested. The models are compared on the basis of their rate of return, maximum drawdown(MDD), and Sharpe ratio. Three underlying funds, AAPL, BCKCX, and BCKCXA, are evaluated under three conditions: with two hidden regimes, with three hidden regimes, and without hidden regimes. For AAPL, the model with three hidden regimes outclasses the other models in terms of rate of return and Sharpe ratio. Similarly, for BCKCX, the model incorporating three hidden regimes surpasses its counterparts in respect to the rate of return and Sharpe ratio. Concerning BCKCXA, the model equipped with three hidden regimes outperforms the other models in terms of rate of return, MDD, and Sharpe ratio. Comparatively, these outcomes emphasize that the integration of hidden regimes can significantly augment the robustness of a model, concurrently enhancing its performance even in the worst-case scenario.

These findings underscore that incorporating the hidden regimes of the financial market

can significantly enhance a model's ability to capture complex patterns in the worst-case scenario. By integrating financial market information, the model's robustness can be enhanced. However, it is essential to note that the optimal number of hidden regimes may vary depending on the specific dataset and problem. Incorporating an excessive number of hidden regimes may cause overfitting, while employing too few may result in a model that oversimplifies, failing to capture significant patterns within the data. Therefore, it is crucial to carefully evaluate the performance of different model architectures and choose the one that best fits the specific task. In this study, we faced computational limitations that prevented us from exploring models with more than three hidden regimes. As such, we only conducted experiments on the models with two and three hidden regimes. Despite computational limitations, comparing the performance of models with two and three hidden regimes effectively illustrates the benefits of integrating market hidden regimes. This integration enhances the model's robustness and performance, particularly under worst-case scenarios.

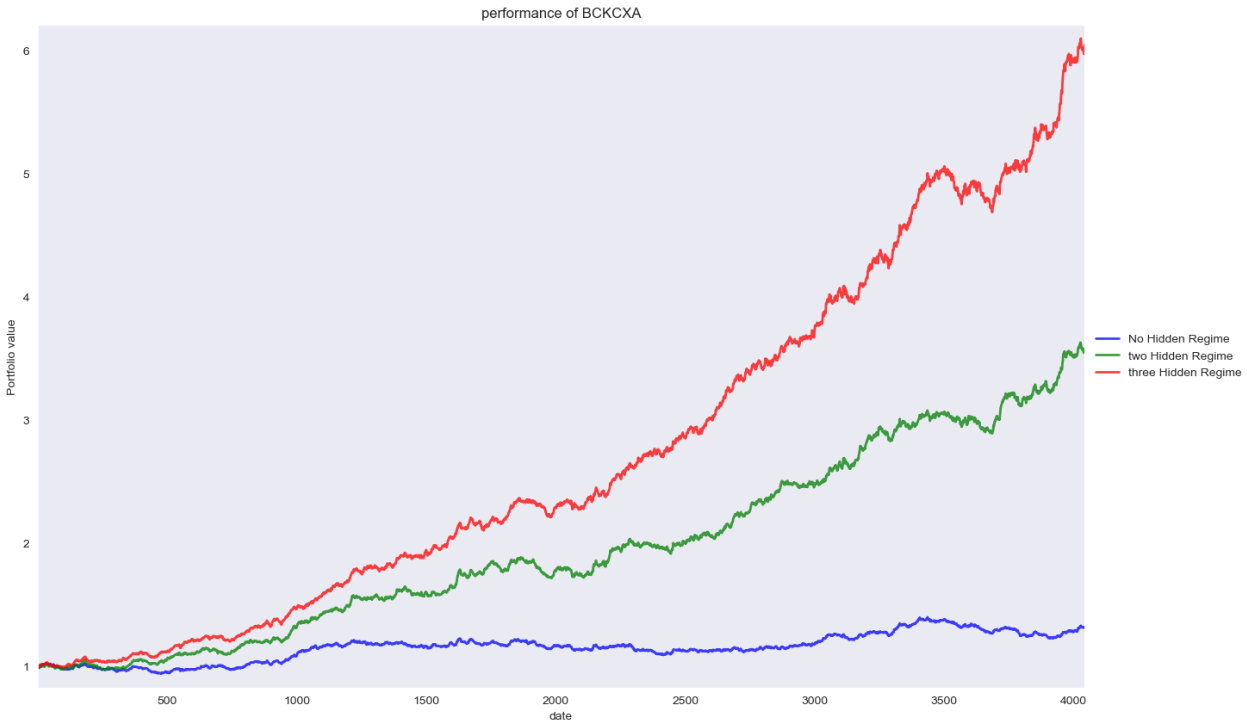


Figure 4.3: The performance of BCKCXA with different hidden regimes

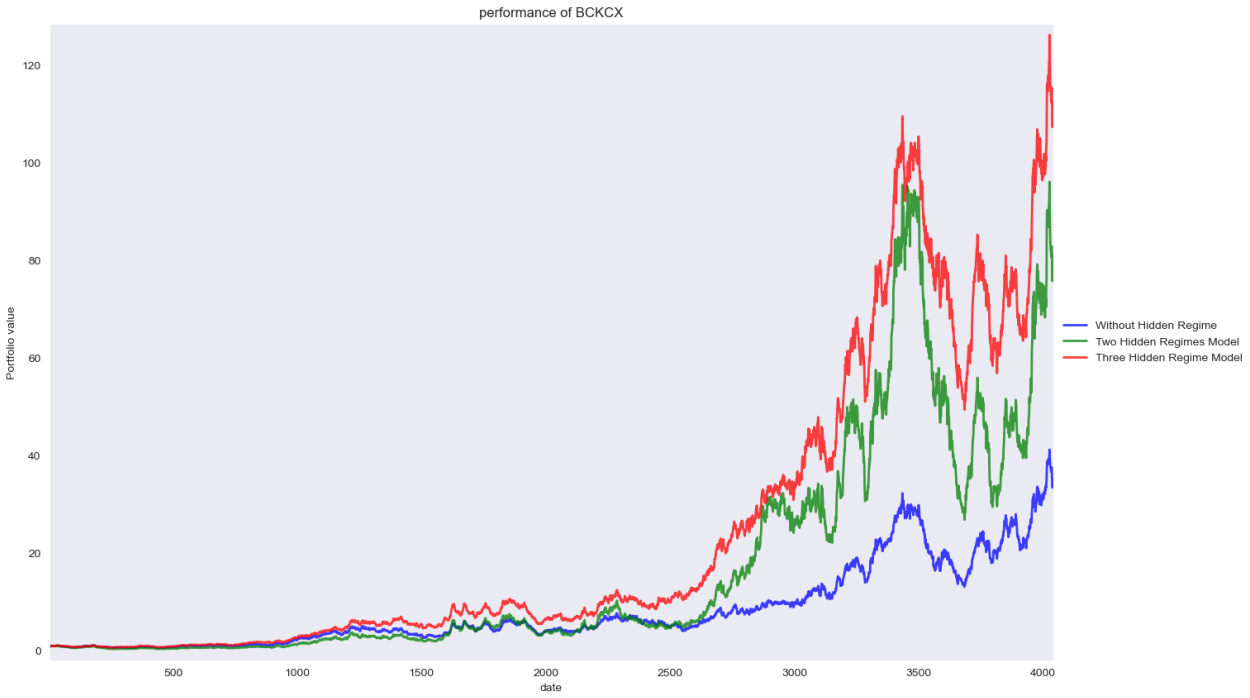


Figure 4.4: The performance of BCKCX with different hidden regimes

Table 4.2: Model Comparison

Fund	**Rate of Return**	**MDD**	**Sharp Ratio**
AAPL (2hr[1])	1532.31%	-12.28%	1.1106
AAPL (3hr[2])	1660.16%	-12.28%	1.1454
AAPL (whr[3])	1629.36%	-12.28%	1.138
BCKCX (2hr[1])	6569.09%	-24.03%	0.7647
BCKCX (3hr[2])	11187.10%	-13.08%	1.0363
BCKCX (whr[3])	1795.12%	-18.95%	0.6614
BCKCXA (2hr[1])	256.99%	-2.64%	1.4619
BCKCXA (3hr[2])	512.75%	-2.44%	2.1063
BCKCXA (whr[3])	32.18%	-2.52%	0.4133

[1] 2hr: two hidden regimes

[2] 3hr: three hidden regimes

[3] whr: without hidden regime

4.7 Conclusion

This study has explored the application of sequential decision models to portfolio management problems, an area characteristically filled with uncertainties owing to the capricious nature of financial markets and macroeconomics. Given the limited information available about immediate rewards and resulting future states for each decision, the accuracy of transition probabilities estimated from financial data cannot always be guaranteed, which may lead to significant deterioration in policy performance in practical scenarios.

In response to these challenges, we proposed an innovative approach that integrates hidden regime rules into a Markov Decision Process (MDP) framework for finite and infinite horizon problems with a finite state and action space. This integration empowers decision-makers to take into account the hidden regimes of the financial market and make more informed decisions in the face of uncertainty. One of the distinguishing features of this setting is its ability to model the uncertainty over transition probabilities, contingent on the current state and the hidden regime of the financial market, further enhancing the robustness and adaptability of our approach.

In this research, we construct a robust dynamic programming framework that incorporates hidden regimes, applicable to both finite and infinite horizon problems. This was approached in two different ways. From the perspective of stochastic control, we demonstrate that fundamental results within dynamic programming (DP) theory are naturally derived when the rectangularity assumption is satisfied. On the other hand, from the standpoint of reinforcement learning, we propose an innovative algorithm. This algorithm takes into account a coherent risk measure along with the risk envelope of transition probabilities that include hidden regimes. The primary aim of this algorithm is to address the portfolio management problem. This dual perspective approach provides a comprehensive understanding of the dynamics and allows us to tackle the problem from multiple angles. It not only shows the robustness of the dynamic programming framework when incorporating hidden regimes, but also illustrates the adaptability and effectiveness of reinforcement learning algorithms in managing portfolio risk in uncertain environments. This study signifies an important step towards more robust and efficient decision-making models in the face of financial market

uncertainties.

However, our work is far from complete. Future work could extend to a continuous (large-scale) state and action space. This would allow our models and algorithms to maintain their robustness and efficiency in larger scale and more complex systems, providing a more comprehensive solution to tackle uncertainties in financial markets. More specifically, the method to maintain and expand the robustness of our dynamic programming framework in large-scale state and action spaces, and how to improve and extend our reinforcement learning algorithms to manage and adapt to risks in various financial environments more effectively can be explored. We look forward to pushing our research to a new level through these extensions and improvements, playing a more significant role in future financial decision-making.

5 Conclusions and Future Work

In this dissertation, we have delved into the complexities of sequential decision-making in the realm of financial management, with a particular focus on portfolio management. We have explored the challenges posed by hidden regimes and the uncertainties inherent in financial markets, and proposed robust and risk-sensitive strategies to navigate these complexities.

We have presented two novel risk-sensitive portfolio management strategies in Chapter 2, utilizing reinforcement learning techniques to construct risk-sensitive policies that protect investors from significant losses. Our empirical experiments on real-world data have demonstrated the effectiveness of these approaches.

Chapter 3 discussed a dynamic scenario of purchasing life insurance and investing in a fluctuating financial market. We suggested robust optimal strategies for life insurance and investment-consumption under regime-switching alpha-ambiguity utility framework, offering a sophisticated approach to wealth and income management over an unpredictable lifetime.

Chapter 4 integrated hidden regimes into the Markov Decision Process, addressing the robust and risk-sensitive MDP problem from both reinforcement learning and stochastic control perspectives. We modeled uncertainty over transition probabilities using the current state and the financial market's hidden regimes, constructing robust and risk-sensitive policies from two different ways.

Through this research, we have made significant contributions to the expanding body of knowledge on financial decision-making, particularly in integrating Markov decision process with hidden regimes. We posit that our work contributes significantly to the contemporary discourse on financial decision-making by providing actionable strategies for navigating the intricate dynamics of financial markets. Future studies will continue to refine our understanding of these complex mechanisms and explore additional ways to enhance robustness

in uncertain environments. This includes the development of more sophisticated models for continuous state and action space, the examination of larger datasets to validate our findings, and the application of advanced algorithms to better capture the underlying dynamics of the financial markets. In a world characterized by growing financial complexities and uncertainties, we believe that the evolution of robust decision-making frameworks is critical. Hence, our research agenda will remain focused on exploring and creating novel methodologies that offer improved risk management while maintaining robustness against the capricious nature of financial markets. We look forward to shedding more light on how to better manage risk and make decisions that are robust in the face of uncertainty and ambiguity.

A Appendix

A.1 Convergence Property of Our Distributional DDPG Model

Let us make use of the probability space $(\Omega, \mathcal{F}, P)$ and view value distributions as random vectors with finite moments in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as usual in the distributional RL model. Consider the process $Z_{k+1} := \mathcal{T}^{\mu} Z_k$ starting with some Z_0. The distributional Bellman operator T^{μ} is a contraction mapping whose unique fixed point is the random return Z^{μ} using the 2-Wasserstein distance. Then, the sequence $\{Z_k\}$ converges to Z^{μ} in distribution. However, this does not necessarily mean pointwise convergence of the sequence $\{Z_k\}$ to Z^{μ}. Bellemare et al. (2017) mention that all moments also converge, in particular $\mathbb{E}[Z_k]$ converges, but one cannot directly use the Wasserstein metric to get the variance convergence.

Let $\mathcal{Q}_k$ and $\mathcal{V}_k$ be the mean and variance of Z_k. Then we have the following convergence results for the proposed distributional DDPG model.

Lemma A.1.1 *For the sequences of mean and variance of Z_k, the following inequalities hold for $k = 1, 2, ...$*

$$||\mathcal{Q}_{k+1} - \mathcal{Q}_k||_{\infty} \leq \gamma ||\mathcal{Q}_k - \mathcal{Q}_{k-1}||_{\infty},$$

$$||\mathcal{V}_{k+1} - \mathcal{V}_k||_{\infty} \leq \gamma^2 ||\mathcal{V}_k - \mathcal{V}_{k-1}||_{\infty},$$

where γ is the discount factor.

Proof. Since $\mathcal{Q}_{k+1} = \mathbb{E}[Z_{k+1}] = \mathbb{E}[\mathcal{T}^\mu Z_k]$, we have

$$\begin{aligned}
||\mathcal{Q}_{k+1} - \mathcal{Q}_k||_\infty &= ||\mathbb{E}[\mathcal{T}^\mu Z_k] - \mathbb{E}[\mathcal{T}^\mu Z_{k-1}]||_\infty \\
&= \sup_{s,a} \gamma |\mathbb{E}[P^\mu Z_k(s,a)] - \mathbb{E}[P^\mu Z_{k-1}(s,a)]| \\
&= \sup_{s,a} \gamma |\mathbb{E}[Z_k(s',a')] - \mathbb{E}[Z_{k-1}(s',a')]| \quad (s' \sim p(\cdot|s,a), a' \sim \mu(\cdot|s)) \\
&\leq \sup_{s',a'} \gamma |\mathbb{E}[Z_k(s',a')] - \mathbb{E}[Z_{k-1}(s',a')]| \\
&= \gamma ||\mathbb{E}[Z_k] - \mathbb{E}[Z_{k-1}]||_\infty = \gamma ||\mathcal{Q}_k - \mathcal{Q}_{k-1}||_\infty.
\end{aligned}$$

Also, since $r(s,a)$ and $P^\mu Z_k(s,a)$ are independent, we get

$$\begin{aligned}
||\mathcal{V}_{k+1} - \mathcal{V}_k||_\infty &= ||\mathrm{Var}(\mathcal{T}^\mu Z_k) - \mathrm{Var}(\mathcal{T}^\mu Z_{k-1})||_\infty \\
&= \sup_{s,a} \gamma^2 |\mathrm{Var}(P^\mu Z_k(s,a)) - \mathrm{Var}(P^\mu Z_{k-1}(s,a))| \\
&\leq \sup_{s',a'} \gamma^2 |\mathrm{Var}(Z_k(s',a')) - \mathrm{Var}(Z_{k-1}(s',a'))| \\
&= \gamma^2 ||\mathrm{Var}(Z_k) - \mathrm{Var}(Z_{k-1})||_\infty = \gamma^2 ||\mathcal{V}_k - \mathcal{V}_{k-1}||_\infty.
\end{aligned}$$

This completes the proof. ■

Lemma A.1.2 *With the discount factor $\gamma < 1$, $\{\mathcal{Q}_k\}$ and $\{\mathcal{V}_k\}$ are Cauchy sequences in L^∞.*

Proof. We need to show that for every positive $\epsilon > 0$, there is a positive integer N such that for every $m, n > N$, $||\mathcal{V}_m - \mathcal{V}_n||_\infty < \epsilon$. Without loss of generality, we assume $||\mathcal{V}_1 - \mathcal{V}_0||_\infty \leq 1$.

By Lemma A.1.1, we have

$$||\mathcal{V}_2 - \mathcal{V}_1||_\infty \leq \gamma^2, \ ||\mathcal{V}_3 - \mathcal{V}_2||_\infty \leq \gamma^4, \ \cdots, \ ||\mathcal{V}_{k+1} - \mathcal{V}_k||_\infty \leq \gamma^{2k}.$$

Also, for $m > n$,

$$\begin{aligned}||\mathcal{V}_m - \mathcal{V}_n||_\infty &\leq ||\mathcal{V}_m - \mathcal{V}_{m-1}||_\infty + \cdots + ||\mathcal{V}_{n+1} - \mathcal{V}_n||_\infty \\ &\leq \gamma^{2(m-1)} + \cdots + \gamma^{2n} = \frac{(1-\gamma^{2(m-n)})}{1-\gamma^2}\gamma^{2n} \\ &\leq (\frac{1}{1-\gamma^2})\gamma^{2n}.\end{aligned}$$

Therefore, we can find a large N such that $\gamma^{2N} < \epsilon(1-\gamma^2)$ for every given $\epsilon > 0$. The result for sequence $\{\mathcal{Q}_k\}$ can be obtained by following essentially the same steps. The proof is completed. ■

Theorem A.1.1 *The sequences of* $\{\mathcal{Q}_k\}$ *and* $\{\mathcal{V}_k\}$ *converge pointwise to their limits in the critic network for policy evaluation.*

Proof. By combining the fact that every Cauchy sequence converges in L^∞ to the limit and Lemma A.1.2, we conclude that the limits of $\{\mathcal{Q}_k\}$ and $\{\mathcal{V}_k\}$ exist and the sequences converge in L^∞. This implies that the convergence takes place for all sample transitions by repeated applications of the distributional Bellman operator $\mathcal{T}^\mu$. This completes the proof.

■

A.2 Experimental Results and Parameters Settings

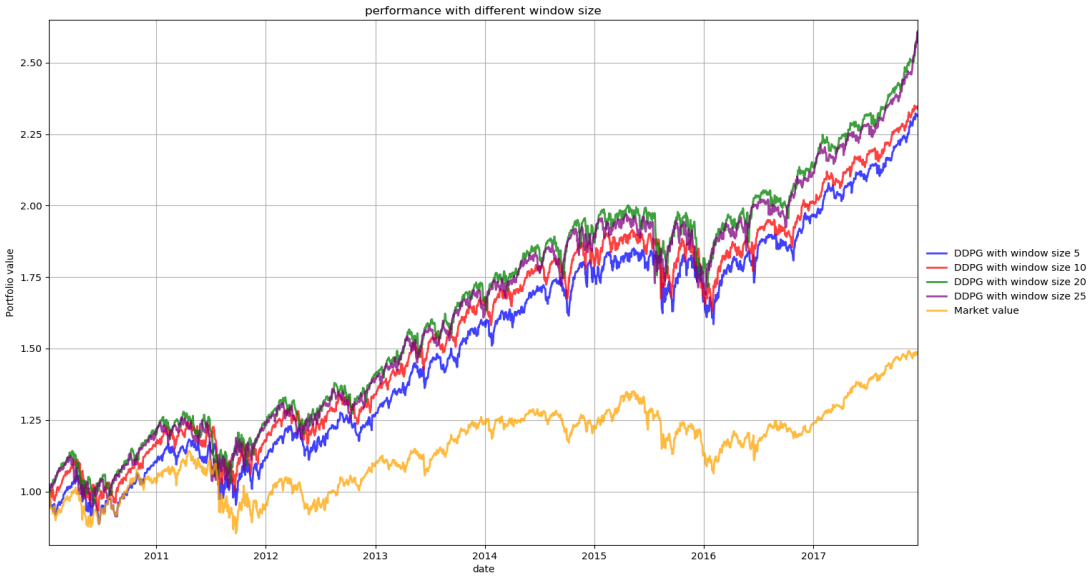


Figure A.1: The portfolio values of classical DDPG with different window sizes during the training period

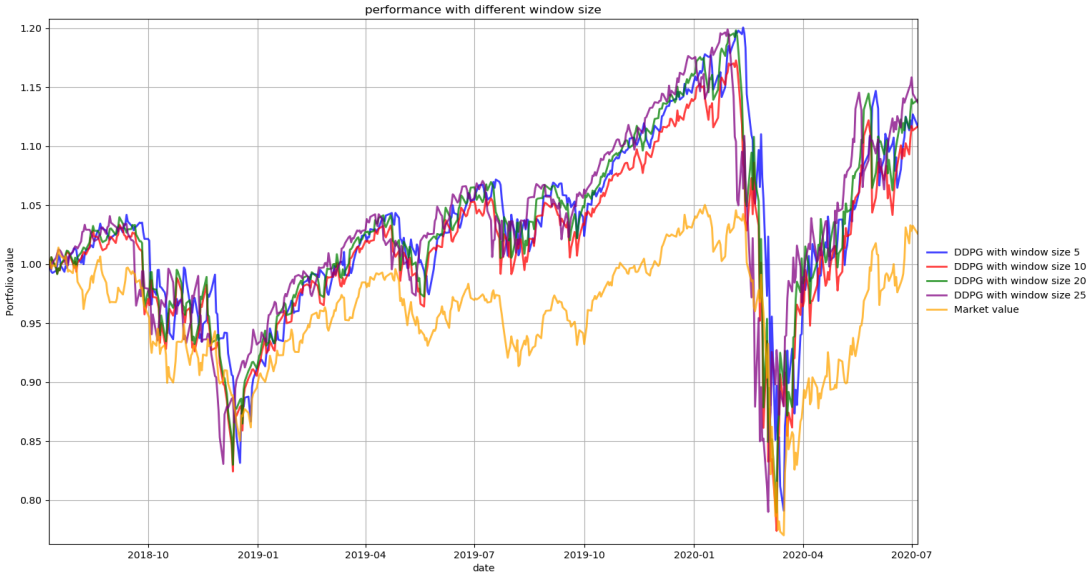


Figure A.2: The portfolio values of classical DDPG with different window sizes during the testing period

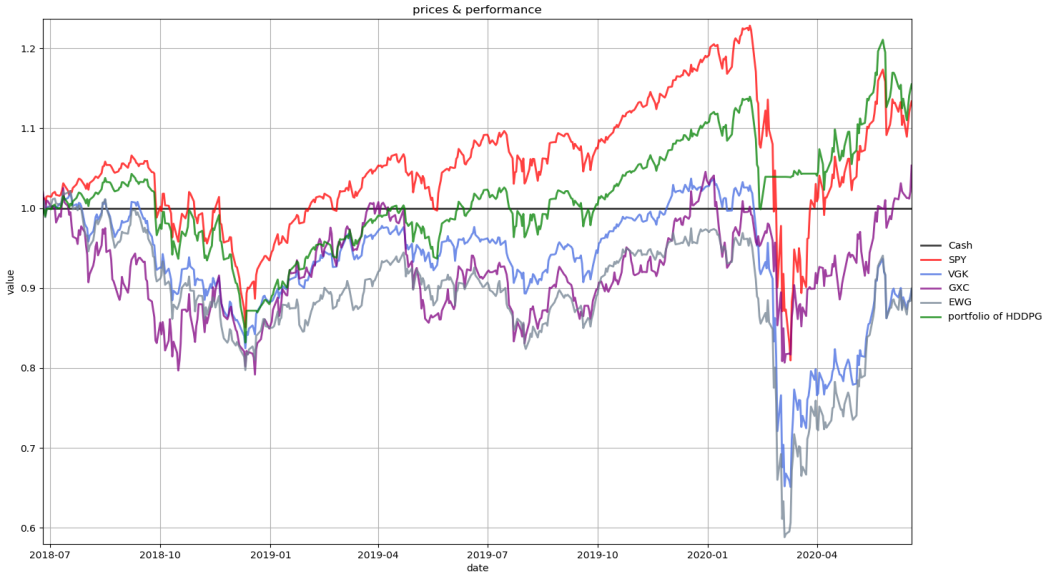


Figure A.3: The price movements of each stock and portfolio values of Hierarchical DDPG with window size of ten-day and CVaR constraint $C = 5\%$.

Table A.1: Hyperparameters of our proposed model

Parameter	DDPG	Distributional DDPG	Hierarchical DDPG
batch size	64	32	64
steps	128	128	128
episode	3000	5000	5000
trading period	1 day	1 day	1 day
learning rate of actor	10^{-5}	10^{-5}	10^{-5}
learning rate of critic	10^{-4}	10^{-4}	10^{-4}
regularization rate	0.001	0.001	0.001
discount rate	0.99	0.99	0.99
memory size	10^6	10^6	10^6
number of layer of actor	5	4	5
number of layer of critic	4	5	4
activition funntion of actor	Relu	Relu	Relu
activition function of critic	Relu	Relu, Softplus	Relu
training set portion	0.8	0.8	0.8
test set portion	0.2	0.2	0.2
commision rate	0.25%	0.25%	0.25%

See Figures A.1-A.3 and Table A.1.

B Appendix

B.1 Proof of Theorem 3.3.1

Proof. Denote by $(\underline{\theta}^*, \underline{\phi_{ij}^*})$ and $(\overline{\theta}^*, \overline{\phi_{ij}^*})$ the candidate distortion processes corresponding to the worst-case measure and candidate best-case measure, which are determined by the inner infimum and supremum parts in the HJB equation (3.12), respectively. By using the first-order condition to equation (3.12) with respect to θ and ϕ_{ij} for the worst-case scenario and the best-case scenario, we obtain the following equations:

$$\pi^2\sigma_i^2 w^2 V_w + \frac{\theta\sigma_i\pi w}{\Psi(t,w,e_i)} = 0, \tag{B.1}$$

$$\pi^2\sigma_i^2 w^2 V_w - \frac{\theta\sigma_i\pi w}{\Psi(t,w,e_i)} = 0,$$

$$\frac{q_{ij}\log(\phi_{ij})}{\Psi(t,w,e_i)} + q_{ij}\big[V(t,w,e_j) - V(t,w,e_i)\big] = 0,$$

$$\frac{q_{ij}\log(\phi_{ij})}{\Psi(t,w,e_i)} - q_{ij}\big[V(t,w,e_j) - V(t,w,e_i)\big] = 0. \tag{B.2}$$

Plugging (3.13) and solving (B.1) and (B.2) lead to

$$\begin{cases} \underline{\theta}^*(t) = -\frac{\mu_i - r}{\sigma_i}\frac{\beta_i}{\gamma - [\hat{\alpha}_i - \alpha]\beta_i}, \\ \underline{\phi_{ij}^*}(t) = \exp\left\{\frac{\beta_i}{1-\gamma}\big[1 - \frac{f(t,e_j)}{f(t,e_i)}\big]\right\}, \end{cases}$$

and

$$\begin{cases} \overline{\theta}^*(t) = \frac{\mu_i - r}{\sigma_i} \frac{\beta_i}{\gamma - [\hat{\alpha}_i - \alpha]\beta_i}, \\ \overline{\phi_{ij}^*}(t) = \exp\left\{ -\frac{\beta_i}{1-\gamma}\left[1 - \frac{f(t,e_j)}{f(t,e_i)}\right] \right\}. \end{cases}$$

Similarly, the robust optimal strategy can be also determined by applying the first-order condition, and is given by

$$\begin{cases} \pi^*(t) = \frac{\mu_i - r}{\sigma_i^2} \frac{-V_w}{w[V_{ww} + [\hat{\alpha}_i - \alpha_i]\Psi V_w^2]}, \\ c^*(t) = V_w^{-\frac{1}{\gamma}}, \\ p^*(t) = [\xi_i \lambda(t)]^{\frac{1}{\gamma}} V_w^{-\frac{1}{\gamma}} \eta^{1-\frac{1}{\gamma}}(t) - w\eta(t). \end{cases} \tag{B.3}$$

Substituting the value function (3.13) into (B.3), we obtain the robust optimal strategy as follows:

$$\begin{cases} \pi^*(t) = \frac{\mu_i - r}{\sigma_i^2} \frac{w + g(t)}{w\left[\gamma - [\hat{\alpha}_i - \alpha_i]\beta_i\right]}, \\ c^*(t) = f^{-\frac{1}{\gamma}}(t, e_i)[w + g(t)], \\ p^*(t) = [\xi_i \lambda(t)]^{\frac{1}{\gamma}} f^{-\frac{1}{\gamma}}(t, e_i) \eta^{1-\frac{1}{\gamma}}(t)[w + g(t)] - w\eta(t). \end{cases}$$

By substituting the established expressions of $\underline{\theta}^*(t)$, $\overline{\theta}^*(t)$, $\underline{\phi_{ij}^*}(t)$, $\overline{\phi_{ij}^*}(t)$, $\pi^*(t)$, $c^*(t)$, and $p^*(t)$ back to the HJB equation (3.12), we obtain that $f(t, e_i)$ satisfies the ODE system (3.17), for each $e_i \in E$ and $g(t)$ is given by (3.18). This completes the proof. ■

B.2 Proof of Theorem 3.3.2

Proof. To prove that the ODE system (3.17) admits a unique solution $f(t, e_i)$, we first show that the solution $f(t, e_i)$, if exists, is bounded and strictly positive.

Rearranging (3.17) gives

$$\frac{df(t,e_i)}{dt}+\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\eta^{1-\frac{1}{\gamma}}(t)\Big]\gamma f^{1-\frac{1}{\gamma}}(t,e_i)+\frac{1-\gamma}{\beta_i}\Big[\alpha_i\sum_{j=1,j\neq i}^{N}q_{ij}\Big(\underline{\phi^*_{ij}}(t)\log\underline{\phi^*_{ij}}(t)-\underline{\phi^*_{ij}}(t)+1\Big)$$
$$-\hat{\alpha}_i\sum_{j=1,j\neq i}^{N}q_{ij}\Big(\overline{\phi^*_{ij}}(t)\log\overline{\phi^*_{ij}}(t)-\overline{\phi^*_{ij}}(t)+1\Big)\Big]f(t,e_i)+\sum_{j=1,j\neq i}^{N}q_{ij}f(t,e_j)\Big[\alpha_i\underline{\phi^*_{ij}}(t)+\hat{\alpha}_i\overline{\phi^*_{ij}}(t)\Big]$$
$$+q_{ii}f(t,e_i)\Big[\alpha_i\underline{\phi^*_{ii}}(t)+\hat{\alpha}_i\overline{\phi^*_{ii}}(t)\Big]-b_i(t)f(t,e_i)=0. \tag{B.4}$$

Define the function $\mathcal{H}\Big(t,\underline{\phi^*_{ij}}(t),\overline{\phi^*_{ij}}(t)\Big)$ as follows:

$$\mathcal{H}\Big(t,\underline{\phi^*_{ij}}(t),\overline{\phi^*_{ij}}(t)\Big):=-b_i(t)+\frac{1-\gamma}{\beta_i}\Big\{\alpha_i\sum_{j=1,j\neq i}^{N}q_{ij}\Big(\underline{\phi^*_{ij}}(t)\log\underline{\phi^*_{ij}}(t)-\underline{\phi^*_{ij}}(t)+1\Big)$$
$$-\hat{\alpha}_i\sum_{j=1,j\neq i}^{N}q_{ij}\Big(\overline{\phi^*_{ij}}(t)\log\overline{\phi^*_{ij}}(t)-\overline{\phi^*_{ij}}(t)+1\Big)\Big\}+q_{ii}\Big[\alpha_i\underline{\phi^*_{ii}}(t)+\hat{\alpha}_i\overline{\phi^*_{ii}}(t)\Big].$$

Equation (B.4) can be rewritten as:

$$\frac{df(t,e_i)}{dt}+\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\eta^{1-\frac{1}{\gamma}}(t)\Big]\gamma f^{1-\frac{1}{\gamma}}(t,e_i)+\sum_{j=1,j\neq i}^{N}q_{ij}f(t,e_j)\Big[\alpha_i\underline{\phi^*_{ij}}(t)+\hat{\alpha}_i\overline{\phi^*_{ij}}(t)\Big]$$
$$+\mathcal{H}\Big(t,\underline{\phi^*_{ij}}(t),\overline{\phi^*_{ij}}(t)\Big)f(t,e_i)=0, \tag{B.5}$$

which immediately leads to the following representation:

$$f(t,e_i)=e^{\int_t^T\mathcal{H}\left(u,\underline{\phi^*_{ij}}(u),\overline{\phi^*_{ij}}(u)\right)du}+\int_t^T e^{\int_t^s\mathcal{H}\left(u,\underline{\phi^*_{ij}}(u),\overline{\phi^*_{ij}}(u)\right)du}\Bigg\{\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(s)\eta^{1-\frac{1}{\gamma}}(s)\Big]\gamma f^{1-\frac{1}{\gamma}}(s,e_i)$$
$$+\sum_{j=1,j\neq i}^{N}q_{ij}f(s,e_j)\Big[\alpha_i\underline{\phi^*_{ij}}(s)+\hat{\alpha}_i\overline{\phi^*_{ij}}(s)\Big]\Bigg\}ds. \tag{B.6}$$

Define $t_0^i:=\sup\{t\in[0,T]|f(t,e_i)\leq 0\}$, for all $i=1,2,\ldots,N$, and $t_0:=t_0^1\vee t_0^2\vee\cdots\vee t_0^N$. As we know $\sup\emptyset=-\infty$, then the range of t_0 is $\{-\infty\}\cup[0,T]$. From $f(T,e_i)=\zeta_i>0$, and the definition of t_0, we have that $f(v,e_i)>0$, for each $v\in(t_0,T]$ and $i=1,2,\ldots,N$.

Furthermore since $q_{ij} > 0$, for $j \neq i$, $\gamma > 0$, $\underline{\phi^*_{ij}}(t) \geq 0$ and $\overline{\phi^*_{ij}}(t) \geq 0$ for any $t \in [0, T]$, respectively, we have

$$\Big[1 + \xi_i^{\frac{1}{\gamma}} \lambda^{\frac{1}{\gamma}}(s) \eta^{1-\frac{1}{\gamma}}(s)\Big] \gamma f^{1-\frac{1}{\gamma}}(s, e_i) + \sum_{j=1, j\neq i}^{N} q_{ij} f(s, e_j) \Big[\alpha_i \underline{\phi^*_{ij}}(s) + \hat{\alpha}_i \overline{\phi^*_{ij}}(s)\Big] > 0, \tag{B.7}$$

for any $s \in (t_0, T]$.

On the other hand, from the expressions of $\underline{\phi^*_{ij}}(t)$ and $\overline{\phi^*_{ij}}(t)$ in Theorem 3.3.1, we have that for any $t \in (t_0, T]$,

$$\begin{cases} 0 \leq \underline{\phi^*_{ij}}(t) \leq e^{\frac{\beta_i}{1-\gamma}}, & \text{for } 0 < \gamma < 1, \\ 0 \leq \underline{\phi^*_{ij}}(t) \leq e^{-\frac{\beta_i}{1-\gamma} \frac{f(t,e_j)}{f(t,e_i)}}, & \text{for } \gamma > 1, \end{cases}$$

and

$$\begin{cases} 0 \leq \overline{\phi^*_{ij}}(t) \leq e^{\frac{\beta_i}{1-\gamma} \frac{f(t,e_j)}{f(t,e_i)}}, & \text{for } 0 < \gamma < 1, \\ 0 \leq \overline{\phi^*_{ij}}(t) \leq e^{-\frac{\beta_i}{1-\gamma}}, & \text{for } \gamma > 1. \end{cases}$$

Thus, we can see that $\underline{\phi^*_{ij}}(t)$ and $\overline{\phi^*_{ij}}(t)$ are bounded if we can show the function $f(t, e_i)$ is bounded from above for any $t \in (t_0, T]$ and $i = 1, 2, ..., N$, which will verified momentarily. Moreover, it is obvious that $\phi^*_{ij}(t) \log \phi^*_{ij}(t) \to 0$ when $\phi^*_{ij}(t) \to 0$. Then, combining the boundedness of market coefficients and the above properties of $\underline{\phi^*_{ij}}(t)$ and $\overline{\phi^*_{ij}}(t)$ guarantees that the continuous function $\mathcal{H}\big(t, \underline{\phi^*_{ij}}(t), \overline{\phi^*_{ij}}(t)\big)$ is bounded, for any $t \in (t_0, T]$. Therefore, there must exist a lower bound H such that $\mathcal{H}\big(t, \underline{\phi^*_{ij}}(t), \overline{\phi^*_{ij}}(t)\big) \geq H > -\infty$, for any $t \in (t_0, T]$.

Setting $t = t_0$ on both sides of equation (B.6) yields that

$$\begin{aligned} 0 \geq f(t_0, e_i) = e^{\int_{t_0}^{T} \mathcal{H}(u, \underline{\phi^*_{ij}}(u), \overline{\phi^*_{ij}}(u)) du} + \int_{t_0}^{T} e^{\int_{t_0}^{s} \mathcal{H}(u, \underline{\phi^*_{ij}}(u), \overline{\phi^*_{ij}}(u)) du} \Bigg\{ \Big[1 + \xi_i^{\frac{1}{\gamma}} \lambda^{\frac{1}{\gamma}}(s) \eta^{1-\frac{1}{\gamma}}(s)\Big] \gamma f^{1-\frac{1}{\gamma}}(s, e_i) \\ + \sum_{j=1, j\neq i}^{N} q_{ij} f(s, e_j) \Big[\alpha_i \underline{\phi^*_{ij}}(s) + \hat{\alpha}_i \overline{\phi^*_{ij}}(s)\Big] \Bigg\} ds > 0. \end{aligned}$$

This is a contradiction. Therefore, $f(t, e_i)$ is positive, for any $t \in [0, T]$ and $i = 1, 2, ..., N$.

Next, we show $f(t,e_i)$ is strictly positive and bounded. We first find a lower bound $\underline{f}>0$ such that $f(t,e_i)\geq \underline{f}$. Since

$$\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(t)\eta^{1-\frac{1}{\gamma}}(t)\Big]\gamma f^{1-\frac{1}{\gamma}}(t,e_i)+\sum_{j=1,j\neq i}^{N}q_{ij}f(t,e_j)\Big[\alpha_i\underline{\phi_{ij}^*}(t)+\hat{\alpha}_i\overline{\phi_{ij}^*}(t)\Big]>0,$$

we have

$$\begin{aligned}f(t,e_i)=e^{\int_t^T\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}+\int_t^T e^{\int_t^s\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}\Bigg\{\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(s)\eta^{1-\frac{1}{\gamma}}(s)\Big]\gamma f^{1-\frac{1}{\gamma}}(s,e_i)\\+\sum_{j=1,j\neq i}^{N}q_{ij}f(s,e_j)\Big[\alpha_i\underline{\phi_{ij}^*}(s)+\hat{\alpha}_i\overline{\phi_{ij}^*}(s)\Big]\Bigg\}ds>e^{\int_t^T\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}.\end{aligned}$$

As derived previously, $\underline{\phi_{ij}^*}(t)$, $\overline{\phi_{ij}^*}(t)$, and $\mathcal{H}\big(t,\underline{\phi_{ij}^*}(t),\overline{\phi_{ij}^*}(t)\big)$ are always bounded for any $t\in[0,T]$. Therefore, we can find a positive constant $\underline{f}$ such that $f(t,e_i)\geq\underline{f}$, that is, $f(t,e_i)$ is strictly positive.

Now, we are in a position to show $f(t,e_i)$ is bounded from above. To that end, we derive

$$\begin{aligned}f(t,e_i)&=e^{\int_t^T\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}+\int_t^T e^{\int_t^s\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}\Bigg\{\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(s)\eta^{1-\frac{1}{\gamma}}(s)\Big]\gamma f^{1-\frac{1}{\gamma}}(s,e_i)\\&\quad+\sum_{j=1,j\neq i}^{N}q_{ij}f(s,e_j)\Big[\alpha(s)\underline{\phi_{ij}^*}(s)+\hat{\alpha}(s)\overline{\phi_{ij}^*}(s)\Big]\Bigg\}ds\\&\leq e^{\int_t^T\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}+\int_t^T e^{\int_t^s\mathcal{H}\left(u,\underline{\phi_{ij}^*}(u),\overline{\phi_{ij}^*}(u)\right)du}\Bigg\{C\Big[1+\xi_i^{\frac{1}{\gamma}}\lambda^{\frac{1}{\gamma}}(s)\eta^{1-\frac{1}{\gamma}}(s)\Big]\gamma f(s,e_i)\\&\quad+\sum_{j=1,j\neq i}^{N}q_{ij}f(s,e_j)\Big[\alpha(s)\underline{\phi_{ij}^*}(s)+\hat{\alpha}(s)\overline{\phi_{ij}^*}(s)\Big]\Bigg\}ds\\&\leq C\Bigg\{1+\sum_{j=1}^{N}\int_t^T f(s,e_j)ds\Bigg\},\end{aligned}\tag{B.8}$$

where C is a positive constant, depending on the lower bound $\underline{f}$, and varies from line to line.

Thus, summing up (B.8) from $i=1$ to $i=N$ and denoting $f^S(t):=\sum_{i=1}^N f(t,e_i)$, we have

$$f^S(t)\leq C\cdot N\Bigg\{1+\int_t^T f^S(s)ds\Bigg\}.$$

An application of Grönwall's inequality yields an upper bound $\overline{f}$ on $f^S(t)$, that is, $f^S(t) \leq \overline{f}$ for each $t \in \mathcal{T}$. Since $f(t, e_i) > 0$ for each $e_i \in E$, $f(t, e_i)$ is also bounded for each $e_i \in E$ and $t \in \mathcal{T}$. Indeed,

$$f(t, e_i) = f^S(t) - \sum_{j=1, j\neq i}^{N} f^S(t, e_j) \leq \overline{f}.$$

Combining the above derivation, we can conclude that there exists two positive constants $\underline{f}$ and $\overline{f}$ such that $f(t, e_i) \in [\underline{f}, \overline{f}]$ and $t \in \mathcal{T}$. It suffices to consider the existence and uniqueness of a solution $f(t, e_i)$ within the established lower bound and upper bound.

For that purpose, we consider the Banach space $C([0, T]; [\underline{f}, \overline{f}])$ and rewrite equation (3.17) as

$$\frac{df(t, e_i)}{dt} = \Gamma(t, f),$$

where

$$\begin{aligned}\Gamma(t, f) := &- \Big[1 + \xi_i^{\frac{1}{\gamma}} \lambda^{\frac{1}{\gamma}}(t) \eta^{1-\frac{1}{\gamma}}(t)\Big] \gamma f^{1-\frac{1}{\gamma}}(t, e_i) - \mathcal{H}\Big(t, \underline{\phi^*_{ij}}(t), \overline{\phi^*_{ij}}(t)\Big) f(t, e_i) \\ &- \sum_{j=1, j\neq i}^{N} q_{ij} f(t, e_j) \Big[\alpha_i \underline{\phi^*_{ij}}(t) + \hat{\alpha}_i \overline{\phi^*_{ij}}(t)\Big].\end{aligned}$$

Taking the first-order derivatives of $\Gamma(t, f)$ with respect to f, we get:

$$\begin{aligned}\frac{\partial}{\partial f}\Gamma(t, f) = &- \frac{1-\gamma}{\beta_i} \left\{ \alpha_i \Bigg[\sum_{j=1, j\neq i}^{N} q_{ij} \Big(\underline{\phi^*_{ij}}(t) \log \underline{\phi^*_{ij}}(t) - \underline{\phi^*_{ij}}(t) + 1 \Big) + \sum_{j=1, j\neq i}^{N} q_{ij} \log \underline{\phi^*_{ij}}(t) \frac{\partial \underline{\phi^*_{ij}}(t)}{\partial f} \Bigg] \right\} \\ &+ \frac{1-\gamma}{\beta_i} \left\{ \hat{\alpha}_i \Bigg[\sum_{j=1, j\neq i}^{N} q_{ij} \Big(\overline{\phi^*_{ij}}(t) \log \overline{\phi^*_{ij}}(t) - \overline{\phi^*_{ij}}(t) + 1 \Big) + \sum_{j=1, j\neq i}^{N} q_{ij} \log \overline{\phi^*_{ij}}(t) \frac{\partial \overline{\phi^*_{ij}}(t)}{\partial f} \Bigg] \right\} \\ &- \left\{ \Big(\alpha_i \frac{\partial \underline{\phi^*_{ij}}(t)}{\partial f} + \hat{\alpha}_i \frac{\partial \overline{\phi^*_{ij}}(t)}{\partial f} \Big) \sum_{j=1, j\neq i}^{N} q_{ij} f(t, e_j) + q_{ii} \Big[\alpha_i \underline{\phi^*_{ii}}(t) + \hat{\alpha}_i \overline{\phi^*_{ii}}(t) \Big] \right\} + b_i(t) \\ &+ \Big[1 + \xi_i^{\frac{1}{\gamma}} \lambda^{\frac{1}{\gamma}}(t) \eta^{1-\frac{1}{\gamma}}(t)\Big] (1-\gamma) f^{-\frac{1}{\gamma}}(t, e_i),\end{aligned}$$

where

$$\frac{\partial \underline{\phi^*_{ij}}(t)}{\partial f} = \underline{\phi^*_{ij}}(t) \times \frac{\beta_i}{(1-\gamma)} \frac{f(t, e_j)}{f^2(t, e_i)} \quad \text{and} \quad \frac{\partial \overline{\phi^*_{ij}}(t)}{\partial f} = -\overline{\phi^*_{ij}}(t) \times \frac{\beta_i}{(1-\gamma)} \frac{f(t, e_j)}{f^2(t, e_i)}.$$

From the above equations, we obtain that the function $\Gamma(t,f)$ is differentiable and continuous with respect to f. Since the market coefficients, $\underline{\phi_{ij}^*}(t)$, $\overline{\phi_{ij}^*}(t)$, and $f(t,e_i)$ are bounded, the partial derivative $\frac{\partial}{\partial f}\Gamma(t,f)$ is also bounded. Thus, $\Gamma(t,f)$ satisfies the Lipschitz condition over $[\underline{f},\overline{f}]$ for $t\in\mathcal{T}$. Applying the mean-value theorem for any $f_1,f_2\in[\underline{f},\overline{f}]$, we obtain that there exists a constant L such that

$$|\Gamma(t,f_1)-\Gamma(t,f_2)|\leq L|f_1-f_2|.$$

Using the Banach fixed point theorem, we have that (B.5) admits a unique solution on $C([0,T];[\underline{f},\overline{f}])$. This completes the proof. ■

B.3 Proof of Theorem 3.3.3

Proof. We define the cost function $\mathcal{C}(s,t;\pi,c,p;\theta,\phi)$ as

$$\begin{aligned}\mathcal{C}(s,t;\pi,c,p;\theta,\phi) :=\ & e^{-\int_t^s\delta(v)dv}\bar{F}(s,t)V(s,W(s),\mathcal{X}(s))\\ &+\left\{\int_t^s\bar{F}(u,t)\alpha(u)e^{-\int_t^u\delta(v)dv}\Big[U(c(u))+\psi(u)+\lambda(u)\xi(u)U(Z(u))\Big]du\right\}\\ &+\left\{\int_t^s\bar{F}(u,t)\hat{\alpha}(u)e^{-\int_t^u\delta(v)dv}\Big[U(c(u))-\psi(u)+\lambda(u)\xi(u)U(Z(u))\Big]du\right\},\end{aligned}$$

where $V(t,w,e_i)$ satisfies the HJB equation (3.12). Applying Itô's formula to $\mathcal{C}(s,t;\pi,c,p;\theta,\phi)$ with respect to s, we derive that

$$\begin{aligned}d\mathcal{C}(s,t;\pi,c,p;\theta,\phi) =& e^{-\int_t^s\delta(v)dv}\bar{F}(s,t)\big[\mathcal{M}_1^{(\pi,c,p;\theta,\phi)}(s,t)+\mathcal{M}_2^{(\pi,c,p;\theta,\phi)}(s,t)\big]ds\\ &+e^{-\int_t^s\delta(v)dv}\bar{F}(s,t)d\mathscr{L}(s),\end{aligned}$$

where

$$\mathcal{M}_1^{(\pi,c,p;\theta,\phi)}(s,t):=\alpha(s)\Big[\mathcal{L}^{(\pi,c,p;\theta,\phi)}V(s,W(s),\mathcal{X}(s))+U(c(s))+\psi(s)+\lambda(s)\xi(s)U(Z(s))\Big],$$

$$\mathcal{M}_2^{(\pi,c,p;\theta,\phi)}(s,t) := \hat{\alpha}(s)\Big[\mathcal{L}^{(\pi,c,p;\theta,\phi)}V(s,W(s),\mathcal{X}(s)) + U(c(s)) - \psi(s) + \lambda(s)\xi(s)U(Z(s))\Big],$$

and $\{\mathscr{L}(s)|s \in \mathcal{T}\}$ is a local martingale (due to the regularity condition 2) satisfying

$$\begin{aligned} d\mathscr{L}(s) =& V_w(s,W(s),\mathcal{X}(s))\pi(s)\sigma(s)W(s)dB^{\mathbb{Q}}(s) + \sum_{j=1}^{N}\Big[V(s,W(s),e_j) \\ & - V(s,W(s),\mathcal{X}(s-))\Big]d\mathscr{K}_j^{\mathbb{Q}}(s). \end{aligned}$$

Here, for each $j = 1, 2, \cdots, N$, the process $\mathscr{K}_j^{\mathbb{Q}}$ is a compensated counting process for the number of jumps of the chain α from all other states into state e_j, and with a $\mathbb{Q}$-compensator associated with $\mathcal{Q}^{\phi}$ that is the transition matrix of the chain α under $\mathbb{Q}$.

Therefore, there exists a localizing sequence $\{k_n|n = 1, 2, ...\}$ such that the local martingale $\{\mathscr{L}(s)|s \in \mathcal{T}\}$ becomes a true martingale. Taking $s = T$, integrating from t to $T \wedge k_n$, and conditioning both sides of the above equation on $\mathcal{F}_t$ under $\mathbb{Q}$ defined by

$$\begin{aligned} \frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{F}_T} =& \exp\Big\{\int_0^T \theta(t)dB(t) - \frac{1}{2}\int_0^T (\theta(t))^2 dt\Big\} \times \exp\Big\{\sum_{i,j=1,i\neq j}^{N}\int_0^T \log\phi_{ij}(t)dM_{ij}(t) \\ & + \sum_{i,j=1,i\neq j}^{N}\int_0^T \Big[\phi_{ij}(t)\log\phi_{ij}(t) - \phi_{ij}(t) + 1\Big]q_{ij}\mathbb{1}_{\{\mathcal{X}(t^-)=e_i\}}dt\Big\}, \end{aligned}$$

where

$$\theta(t) := \alpha(t)\underline{\theta}(t) + \hat{\alpha}(t)\overline{\theta}(t) \tag{B.9}$$

and

$$\phi_{ij}(t) := \alpha(t)\underline{\phi}_{ij}(t) + \hat{\alpha}(t)\overline{\phi}_{ij}(t), \tag{B.10}$$

we have

$$\begin{aligned} \mathbb{E}_{t,w,i}^{\mathbb{Q}}\big[\mathcal{C}(s\wedge k_n, t;\pi,c,p;\theta,\phi)\big] =& V(t,w,e_i) + \mathbb{E}_{t,w,i}^{\mathbb{Q}}\Big[\int_t^{s\wedge k_n} e^{-\int_t^u \delta(v)dv}\bar{F}(u,t)[\mathcal{M}_1^{(\pi,c,p;\underline{\theta},\underline{\phi})}(u,t) \\ & + \mathcal{M}_2^{(\pi,c,p;\overline{\theta},\overline{\phi})}(u,t)]du\Big]. \end{aligned}$$

Note that if $\alpha(u) = 1$ and $\hat{\alpha}(u) = 0$, then $\mathcal{M}_2^{(\pi,c,p;\underline{\theta},\underline{\phi})}(u) = 0$; otherwise, if $\alpha(u) = 0$ and $\hat{\alpha}(u) = 1$, then $\mathcal{M}_1^{(\pi,c,p;\overline{\theta},\overline{\phi})}(u) = 0$. According to Theorem 3.3.1, the function $\mathcal{M}_1^{(\pi,c,p;\underline{\theta},\underline{\phi})}(u)$ and $\mathcal{M}_2^{(\pi,c,p;\overline{\theta},\overline{\phi})}(s)$ may achieve the maximal or minimal value at zero depending on the value of $\alpha(s)$. The following assertions hold: if $\alpha(u) = 1$, then

(i) $\mathcal{M}_1^{(\pi,c,p;\underline{\theta}^*,\underline{\phi}^*)}(s) \leq 0$, for any $(\pi, c, p) \in \mathcal{A}$, when $(\underline{\theta}^*, \underline{\phi}^*)$ is given by (3.16);

(ii) $\mathcal{M}_1^{(\pi^*,c^*,p^*;\underline{\theta}^*,\underline{\phi}^*)}(s) = 0$, when (π^*, c^*, p^*) and $(\underline{\theta}^*, \underline{\phi}^*)$ are given by (3.14) and (3.16), respectively.

If $\hat{\alpha}(s) = 1$, then

(i') $\mathcal{M}_2^{(\pi,c,p;\overline{\theta}^*,\overline{\phi}^*)}(s) \leq 0$, for any $(\pi, c, p) \in \mathcal{A}$, when $(\overline{\theta}^*, \overline{\phi}^*)$ is given by (3.15);

(ii') $\mathcal{M}_2^{(\pi^*,c^*,p^*;\overline{\theta}^*,\overline{\phi}^*)}(s) = 0$, when (π^*, c^*, p^*) and $(\overline{\theta}^*, \overline{\phi}^*)$ is given by (3.14) and (3.15), respectively.

It follows from Assertions (i) and (i'), respectively, that for any $(\pi, c, p) \in \mathcal{A}$,

$$\begin{aligned}\mathbb{E}_{t,w,i}^{\mathbb{Q}}\Big[\mathcal{C}(T \wedge k_n, t; \pi, c, p; \theta^*, \phi^*)\Big] &= V(t, w, e_i) + \mathbb{E}_{t,w,i}^{\mathbb{Q}}\Bigg[\int_t^{T\wedge k_n} e^{-\int_t^u \delta(v)dv} \bar{F}(u,t)[\mathcal{M}_1^{(\pi,c,p;\underline{\theta}^*,\underline{\phi}^*)}(u,t) \\ &\quad + \mathcal{M}_2^{(\pi,c,p;\overline{\theta}^*,\overline{\phi}^*)}(u,t)]du\Bigg] \\ &\leq V(t, w, e_i),\end{aligned}$$

Since $V(\cdot, \cdot, \mathcal{X}(\cdot))$ is uniformly integrable (Regularity Condition 1), and $\{U(c(t)) + \psi(t) + \lambda(u)U(Z(t))|t \in [0, T]\}$ and $\{U(c(t)) - \psi(t) + \lambda(u)U(Z(t))|t \in [0, T]\}$ are also uniformly integrable (Regularity Conditions 2 and 3), applying Lebesgue's dominated convergence theorem to the conditional expectation on the left-hand side, we get:

$$\mathbb{E}_{t,w,i}^{\mathbb{Q}}\Big[\mathcal{C}(T, t; \pi, c, p; \theta^*, \phi^*)\Big] = \lim_{n\to\infty} \mathbb{E}_{t,w,i}^{\mathbb{Q}}\Big[\mathcal{C}(T \wedge k_n, t; \pi, c, p; \theta^*, \phi^*)\Big] \leq V(t, w, e_i).$$

Therefore, we get:

$$\inf_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \alpha(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))+\bar{F}(u,t)\psi(u)+f(u,t)\xi(u)U(Z(u))\Big]du$$
$$+\alpha(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]+\sup_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \hat{\alpha}(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))$$
$$-\bar{F}(u,t)\psi(u)+f(u,t)\xi(u)U(Z(u))\Big]du+\hat{\alpha}(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]\leq V(t,w,e_i).$$

Taking supremum with respect to $u=(\pi,c,p)\in\mathcal{A}$, we have

$$\sup_{(\pi,c,p)\in\mathcal{A}}\Bigg\{\inf_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \alpha(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))+\bar{F}(u,t)\psi(u)$$
$$+f(u,t)\xi(u)U(Z(u))\Big]du+\alpha(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]$$
$$+\sup_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \hat{\alpha}(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))-\bar{F}(u,t)\psi(u)$$
$$+f(u,t)\xi(u)U(Z(u))\Big]du+\hat{\alpha}(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]\Bigg\}\leq V(t,w,e_i). \tag{B.11}$$

Similarly, we can derive from Assertions (i) and (i') and Assertions (ii) and (ii') that

$$\sup_{(\pi,c,p)\in\mathcal{A}}\Bigg\{\inf_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \alpha(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))+\bar{F}(u,t)\psi(u)$$
$$+f(u,t)\xi(u)U(Z(u))\Big]du+\alpha(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]$$
$$+\sup_{(\theta,\phi)\in\Theta}\mathbb{E}^{\mathbb{Q}}_{t,w,i}\Bigg[\int_t^T \hat{\alpha}(u)e^{-\int_t^u \delta(v)dv}\Big[\bar{F}(u,t)U(c(u))-\bar{F}(u,t)\psi(u)$$
$$+f(u,t)\xi(u)U(Z(u))\Big]du+\hat{\alpha}(T)e^{-\int_t^T \delta(v)dv}\bar{F}(T,t)U(W(T))\Bigg]\Bigg\}=V(t,w,e_i). \tag{B.12}$$

Therefore, combining (B.11) and (B.12), we can conclude that (π^*,c^*,p^*) is the optimal strategy, $(\underline{\theta}^*,\underline{\phi^*})$ and $(\overline{\theta}^*,\overline{\phi^*})$ are the worst-case measure and the best-case measure, respectively. This completes the proof. ■

B.4 Technical Proof for Theorem 3.3.4

Proof. For Condition 1, we only verify that $\{V(t,W(t),\mathcal{X}(t))|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario, and omit the verification of the uniform integrability of $\{V(t,W(t),\mathcal{X}(t))|t\in[0,T]\}$ under the best-case scenario, since it can be proved similarly.

Under the worst-case scenario, substitute (3.14) and (3.15) into (3.6). Then, we have

$$\begin{aligned}\frac{d(W(t)+g(t))}{W(t)+g(t)}=&\Bigg\{\frac{[\mu(t)-r]^2}{\sigma^2(t)[\gamma-(1-2\alpha(t))\beta(t)]}+[r+\eta(t)]-[1+\lambda^{\frac{1}{\gamma}}(t)]f^{-\frac{1}{\gamma}}(t,\mathcal{X}(t))\\&-\frac{[\mu(t)-r]^2\beta(t)}{\sigma^2(t)[\gamma-(1-2\alpha(t))\beta(t)]^2}\Bigg\}dt+\frac{\mu(t)-r}{\sigma(t)[\gamma-(1-2\alpha(t))\beta(t)]}dB^{\mathbb{Q}}(t).\end{aligned}\tag{B.13}$$

It is obvious that (B.13) has a unique solution:

$$\begin{aligned}W(t)+g(t)=&(w_0+g(0))\exp\Bigg\{\int_0^t A^2(s)B(s)+(r+\eta(s))-(1+\lambda^{\frac{1}{\gamma}}(s))f^{-\frac{1}{\gamma}}(s,\mathcal{X}(s))\\&-A^2(s)B^2(s)\beta(s)-\frac{1}{2}A^2(s)B^2(s)ds+\int_0^t A(s)B(s)dB^{\mathbb{Q}}(s)\Bigg\},\end{aligned}\tag{B.14}$$

where

$$A(t)=\frac{\mu(t)-r}{\sigma(t)}\quad\text{and}\quad B(t)=\frac{1}{\gamma-[1-2\alpha(t)]\beta(t)}.$$

Insert (B.14) into the candidate value function (3.13). By the boundedness of $f(t,e_i)$, $g(t)$, and other model parameters, we obtain the following for any $m\geq 1$:

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^m\right]&=\mathbb{E}^{\mathbb{Q}}\left[f^m(t,e_i)\frac{(W(t)+g(t))^{m(1-\gamma)}}{(1-\gamma)^m}\right]\\&\leq K\cdot\mathbb{E}^{\mathbb{Q}}\left[e^{m(1-\gamma)\int_0^t A(s)B(s)dB^{\mathbb{Q}}(s)}\right]\\&=K\cdot\mathbb{E}^{\mathbb{Q}}\Bigg[\underbrace{e^{\frac{1}{2}m^2(1-\gamma)^2\int_0^t A^2(s)B^2(s)ds}}_{\text{bounded}}\\&\quad\times\underbrace{e^{-\frac{1}{2}m^2(1-\gamma)^2\int_0^t A^2(s)B^2(s)ds+m(1-\gamma)\int_0^t A(s)B(s)dB^{\mathbb{Q}}(s)}}_{\text{martingale}}\Bigg]<\infty,\end{aligned}\tag{B.15}$$

where K is a positive constant. The above inequalities in equation (B.15) hold, because $f(s)$, $A(s)$, and $B(s)$ are bounded on $[0,T]$.

Therefore, choosing $m>1$ and taking surpremum of the above equality gives

$$\sup_{t\in[0,T]}\mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^m\right]<\infty.$$

This immediately implies that $\{V(t,W(t),\mathcal{X}(t))|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario. That is, Condition 1 holds.

For Condition 2, we first show under the worst-case scenario

$$\begin{aligned}
I(t) &:= \mathbb{E}^{\mathbb{Q}}_{t,w,i}\left[\exp\left(\frac{1}{2}\int_0^T \underline{\theta^*}(t)^2dt+\sum_{i,j=1,i\neq j}^{N}\int_0^T[\underline{\phi^*_{ij}}(t)\log\underline{\phi^*_{ij}}(t)-\underline{\phi^*_{ij}}(t)+1]q_{ij}dt\right)\right]\\
&= \mathbb{E}^{\mathbb{Q}}_{t,w,i}\left[\exp\left(\frac{1}{2}\int_0^T\frac{[\mu(t)-r]^2\beta^2(t)}{\sigma^2(t)[\gamma-(1-2\alpha(t))\beta(t)]^2}dt\right)\right]\\
&\quad\times\mathbb{E}^{\mathbb{Q}}_{t,w,i}\left[\exp\left(\sum_{i,j=1,i\neq j}^{N}\int_0^T q_{ij}\left[e^{-\frac{\beta(t)}{1-\gamma}\left(1-\frac{f(t,e_j)}{f(t,e_i)}\right)}\left[-\frac{\beta(t)}{1-\gamma}\left(1-\frac{f(t,e_j)}{f(t,e_i)}\right)\right]\right.\right.\right.\\
&\quad\left.\left.\left.-e^{-\frac{\beta(t)}{1-\gamma}\left(1-\frac{f(t,e_j)}{f(t,e_i)}\right)}+1\right]dt\right)\right]
\end{aligned}\tag{B.16}$$

is finite since $f(t,e_i)$ is bounded on $[0,T]$. Thus, we have $I(T)<\infty$.

Next, to verify that the $\{U(c(t))+\psi(t)+\lambda(t)U(Z(t))|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario, we need to show that $\{U(c(t))|t\in[0,T]\}$, $\{\psi(t)|t\in[0,T]\}$, and $\{\lambda(t)U(Z(t))|t\in[0,T]\}$ are uniformly integrable, respectively. First verifying that $\{\psi(t)|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario, we show

$$\sup_{t\in[0,T]}\mathbb{E}^{\mathbb{Q}}\left[|\psi(t)|^2\right]<\infty.$$

Insert (π^*, c^*, p^*) and $(\underline{\theta}^*(t), \underline{\phi}^*_{ij}(t))$ into (3.7). Since (B.16) is finite, we have

$$
\begin{aligned}
&\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\big[|\psi(t)|^2\big] \\
&= \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^2\left|\frac{1-\gamma}{\beta}\underline{\theta}^*(t)^2 + \frac{1-\gamma}{\beta}\sum_{i,j=1,j\neq i}^{N} q_{ij}\Big(\underline{\phi}^*_{ij}(t)\log\underline{\phi}^*_{ij}(t) - \underline{\phi}^*_{ij}(t) + 1\Big)\right|^2\right] \\
&\leq \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^4\right]^{\frac{1}{2}} \\
&\quad\times \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{1-\gamma}{\beta}\underline{\theta}^*(t)^2 + \frac{1-\gamma}{\beta}\sum_{i,j=1,j\neq i}^{N} q_{ij}\Big(\underline{\phi}^*_{ij}(t)\log\underline{\phi}^*_{ij}(t) - \underline{\phi}^*_{ij}(t) + 1\Big)\right|^4\right]^{\frac{1}{2}} < \infty.
\end{aligned}
$$

Next, we verify $\{U(c(t))|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario. Substituting (3.14) and (3.16) into $\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\big[|U(c(t))|^2\big]$, we have

$$
\begin{aligned}
\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\big[|U(c(t))|^2\big] &= \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{\big[f^{-\frac{1}{\gamma}}(t,\mathcal{X}(t))(W(t)+g(t))\big]^{1-\gamma}}{1-\gamma}\right|^2\right] \\
&= \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))f^{-\frac{1}{\gamma}}(t,\mathcal{X}(t))\right|^2\right] \\
&\leq K\cdot\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^2\right] < \infty,
\end{aligned}
$$

since the value function $V(t,W(t),\mathcal{X}(t))$ is uniformly integrable and $f(t,\mathcal{X}(t))$ is bounded on $[0,T]$.

Finally, we show that $\{\lambda(t)U(Z(t))|t\in[0,T]\}$ is uniformly integrable under the worst-case scenario. Following the same procedure, inserting (3.14) and (3.16) into $\sup_{t\in[0,T]} E^{\mathbb{Q}}\big[|\lambda(t)U(Z(t))|^2\big]$, we can find

$$
\begin{aligned}
\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\big[|\lambda(t)U(Z(t))|^2\big] &= \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|\lambda(t)\frac{\big[\lambda^{\frac{1}{\gamma}}(t)f^{-\frac{1}{\gamma}}(t,\mathcal{X}(t))[W(t)+g(t)]\eta^{-\frac{1}{\gamma}}(t)\big]^{1-\gamma}}{1-\gamma}\right|^2\right] \\
&= \sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|\lambda^{\frac{1}{\gamma}}(t)f^{-\frac{1}{\gamma}}(t,\mathcal{X}(t))\eta^{1-\frac{1}{\gamma}}(t)V(t,W(t),\mathcal{X}(t))\right|^2\right] \\
&\leq K\cdot\sup_{t\in[0,T]} \mathbb{E}^{\mathbb{Q}}\left[\left|V(t,W(t),\mathcal{X}(t))\right|^2\right] < \infty,
\end{aligned}
$$

since $\lambda(t)$ and $\eta(t)$ are deterministic and bounded on $[0,T]$. That is, Condition 2 is true.

Condition 3, that is, $\{U(c(t)) - \psi(t) + \lambda(t)U(Z(t))|t \in [0,T]\}$ is uniformly integrable for the best-case scenario, and can be shown similarly, thus it is omitted here.

To show Condition 4, denote by

$$\mathcal{L}(s) := \int_t^s V_w(u, W(u), \mathcal{X}(u))\pi(u)\sigma(u)W(u)dB^{\mathbb{Q}}(u).$$

To prove $\mathcal{L}$ is a local martingale, we only need to show that

$$\mathbb{E}^{\mathbb{Q}}\Big[\int_t^T V_w^2(s, W(s), \mathcal{X}(s))\pi^2(s)\sigma(s)^2W(s)^2ds\Big] < \infty. \tag{B.17}$$

From (3.13), we have

$$V_w^2(u, W(u), \mathcal{X}(u)) = f^2(u, \mathcal{X}(u))(W(u) + g(u))^{-2\gamma}. \tag{B.18}$$

Insert $W(t) + g(t)$ and equation (B.18) under the worst-case scenario into (B.17), we have

$$\mathbb{E}^{\mathbb{Q}}\Big[\int_t^T V_w^2(s, W(s), \mathcal{X}(s))\pi^2(s)\sigma(s)^2W(s)^2ds\Big] = \mathbb{E}^{\mathbb{Q}}\Big[\int_t^T V^2(s, W(s), \mathcal{X}(s))A(s)^2B^2(s)ds\Big] < \infty.$$

Since $V(s, W(s), \mathcal{X}(s))$ is uniformly integrable and $A(s)$, $B(s)$ are bounded on $[0,T]$, the above derivations guarantee that $\mathcal{L}$ is an $(\mathbb{F}, \mathbb{Q})$ local martingale. Similar to the above procedure to prove the martingale of $\mathcal{L}$ under the worst-case scenario, we can show the $\mathcal{L}$ is an $(\mathbb{F}, \mathbb{Q})$ local martingale under the best-case scenario. Thus, Condition 4 holds.

So far, Conditions 1-4 have been verified. Therefore, we can conclude that $V(t, w, e_i)$ defined by (3.13) is indeed the corresponding value function of this problem. This completes the proof. ■

C Appendix

Figure C.1: The closing price of AAPL

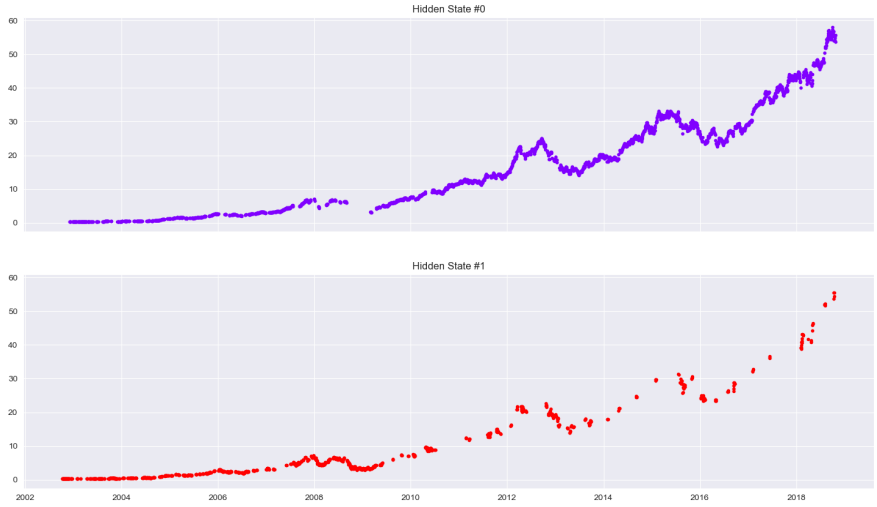


Figure C.2: Two hidden regimes of AAPL

Figure C.3: Three hidden regimes of AAPL

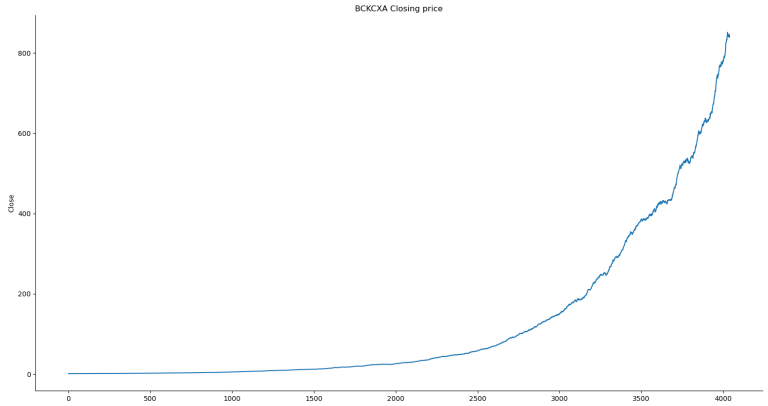


Figure C.4: The closing price of BCKCXA

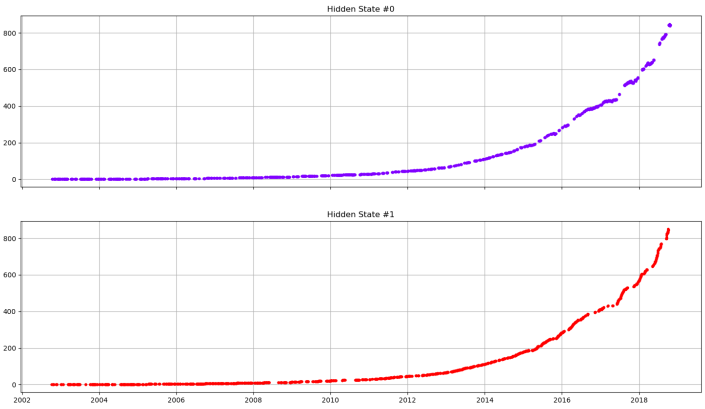


Figure C.5: Two hidden regimes of BCKCXA

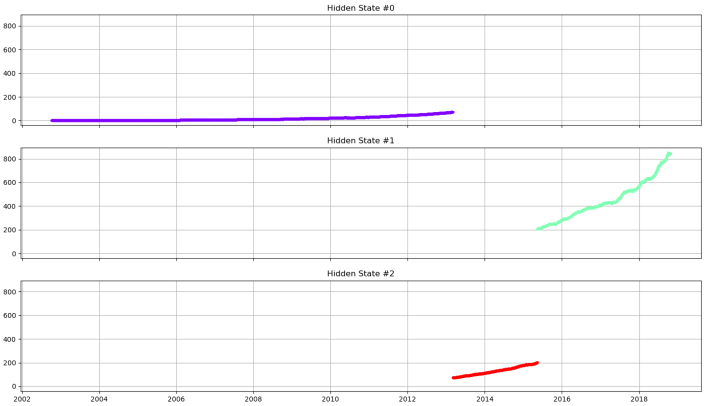


Figure C.6: Three hidden regimes of BCKCXA

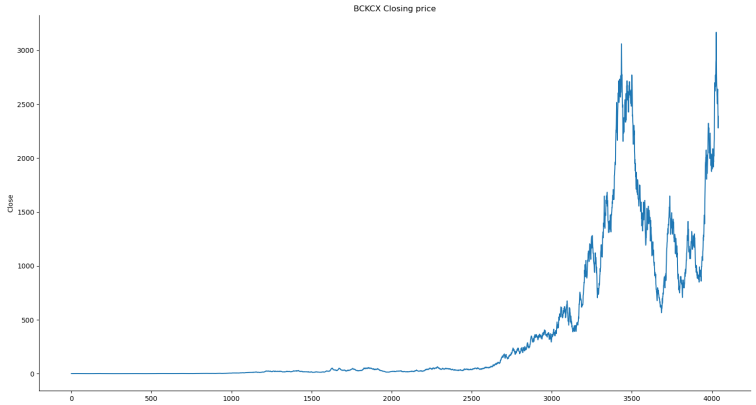


Figure C.7: The closing price of BCKCX

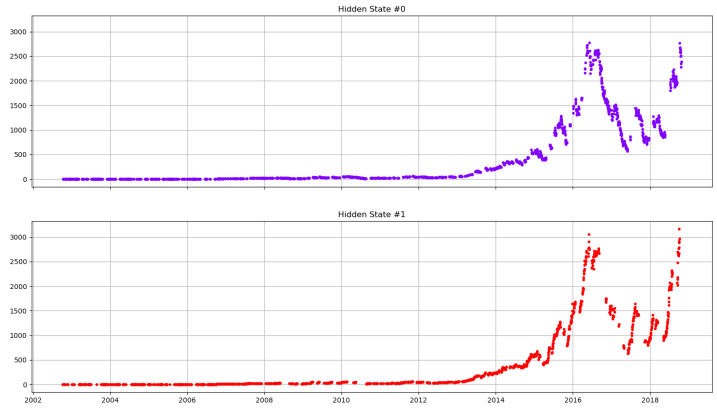


Figure C.8: Two hidden regimes of BCKCX

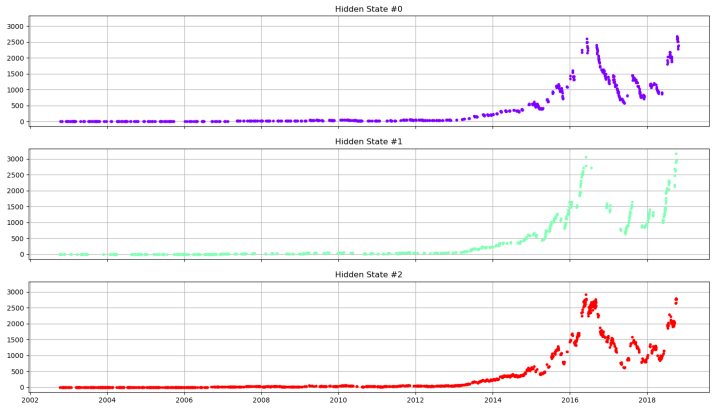


Figure C.9: Three hidden regimes of BCKCX

Bibliography

Abdullah, M. A., Ren, H., Ammar, H. B., Milenkovic, V., Luo, R., Zhang, M., & Wang, J. (2019). Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

Almahdi, S., & Yang, S. Y. (2017). An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, *87*, 267–279.

Anderson, E. W., Hansen, L. P., & Sargent, T. J. (2003). A quartet of semigroups for model specification, robustness, prices of risk, and model detection. *Journal of the European Economic Association*, *1*(1), 68–123.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, *9*(3), 203–228.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., Tb, D., ... Lillicrap, T. (2018). Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, *13*(1), 41–77.

Beissner, P., Lin, Q., & Riedel, F. (2020). Dynamically consistent alpha-maxmin expected utility. *Mathematical Finance*, *30*(3), 1073–1102.

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. *International Conference on Machine Learning*, 449–458.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization* (Vol. 28). Princeton university press.

Ben-Tal, A., & Nemirovski, A. (1998). Robust convex optimization. *Mathematics of operations research*, *23*(4), 769–805.

Bertoluzzo, F., & Corazza, M. (2012). Reinforcement learning for automatic financial trading: Introduction and some applications. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, *33*.

Bertsekas, D., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, 1196–1217.

Bier, V. M., & Connell, B. L. (1994). Ambiguity seeking in multi-attribute decisions: Effects of optimism and message framing. *Journal of Behavioral Decision Making*, *7*(3), 169–182.

Borkar, V., & Jain, R. (2014). Risk-constrained markov decision processes. *IEEE Transactions on Automatic Control*, *59*(9), 2574–2579.

Branger, N., Larsen, L. S., & Munk, C. (2013). Robust portfolio choice with ambiguity and learning about return predictability. *Journal of Banking & Finance*, *37*(5), 1397–1411.

Chen, L., & Gao, Q. (2019). Application of deep reinforcement learning on automated stock trading. *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, 29–33.

Chow, Y., Ghavamzadeh, M., Janson, L., & Pavone, M. (2017). Risk-constrained reinforce-

ment learning with percentile risk criteria. *The Journal of Machine Learning Research*, *18*(1), 6070–6120.

Chow, Y., Tamar, A., Mannor, S., & Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. *arXiv e-prints*, arXiv–1506.

Dayan, P. (2002). Reinforcement learning. *Stevens' Handbook of Experimental Psychology*.

Dearden, R., Friedman, N., & Russell, S. J. (1998). Bayesian q-learning. In *Aaai/iaai*.

De Farias, D. P., & Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations research*, *51*(6), 850–865.

Dietterich, T. G. (1998). The maxq method for hierarchical reinforcement learning. In *Icml* (Vol. 98, pp. 118–126).

Dimmock, S. G., Kouwenberg, R., Mitchell, O. S., & Peijnenburg, K. (2016). Ambiguity aversion and household portfolio choice puzzles: Empirical evidence. *Journal of Financial Economics*, *119*(3), 559–577.

Du, Y., Wang, S., & Huang, L. (2022). Risk-sensitive reinforcement learning: Iterated cvar and the worst path. *arXiv preprint arXiv:2206.02678*.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, *92*(4), 433.

Elliott, R. J., Aggoun, L., & Moore, J. B. (1995). *Hidden markov models: Estimation and control*. Springer, Berlin.

Elliott, R. J., & Siu, T. K. (2009). Robust optimal portfolio choice under Markovian regime-switching model. *Methodology and Computing in Applied Probability*, *11*(2), 145–157.

Engel, Y., Mannor, S., & Meir, R. (2005). Reinforcement learning with gaussian processes. *Proceedings of the 22nd international conference on Machine learning*, 201–208.

Füllbrunn, S., Rau, H. A., & Weitzel, U. (2014). Does ambiguity aversion survive in experimental asset markets? *Journal of Economic Behavior & Organization*, *107*, 810–826.

Gao, Y., Lui, K. Y. C., & Hernandez-Leal, P. (2021). Robust risk-sensitive reinforcement learning agents for trading markets. *arXiv preprint arXiv:2107.08083*.

Gao, Z., Gao, Y., Hu, Y., Jiang, Z., & Su, J. (2020). Application of deep q-network in portfolio management. *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, 268–275.

Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory*, *118*(2), 133–173.

Goyal, V., & Grand-Clement, J. (2023). Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, *48*(1), 203–226.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–84.

Hansen, L. P., & Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*, *91*(2), 60–66.

Hansen, L. P., & Sargent, T. J. (2008). *Robustness*. Princeton university press.

Harnpadungkij, T., Chaisangmongkon, W., & Phunchongharn, P. (2019). Risk-sensitive portfolio management by using distributional reinforcement learning. In *2019 ieee 10th international conference on awareness science and technology (icast)* (pp. 1–6).

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*(1), 5–28.

Hegde, S., Kumar, V., & Singh, A. (2018). Risk aware portfolio construction using deep deterministic policy gradients. In *2018 ieee symposium series on computational intelli-*

gence (ssci) (pp. 1861–1867).

Hou, P., Yeoh, W., & Varakantham, P. (2014). Revisiting risk-sensitive mdps: New algorithms and results. , *24*, 136–144.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, *30*(2), 257–280.

Jeong, G., & Kim, H. Y. (2019). Improving financial trading decisions using deep q-learning: Predicting the number of shares, action strategies, and transfer learning. *Expert Systems with Applications*, *117*, 125–138.

Jiang, Z., & Liang, J. (2017). Cryptocurrency portfolio management with deep reinforcement learning. In *2017 intelligent systems conference (intellisys)* (pp. 905–913).

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206.

Kaufman, D. L., & Schaefer, A. J. (2013). Robust modified policy iteration. *INFORMS Journal on Computing*, *25*(3), 396–410.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, *73*(6), 1849–1892.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2009). Recursive smooth ambiguity preferences. *Journal of Economic Theory*, *144*(3), 930–976.

Krokhmal, P., Palmquist, J., & Uryasev, S. (2002). Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, *4*, 43–68.

Lee, S. I., & Shim, G. (2015). Optimal consumption/investment and life insurance with regime-switching financial market parameters. *Journal of the Korean Society for Industrial and Applied Mathematics*, *19*(4), 429–441.

Levy, K., Vazquez-Abad, F. J., & Costa, A. (2006). Adaptive stepsize selection for online q-learning in a non-stationary environment. , 372–377.

Li, B., Li, D., & Xiong, D. (2016). Alpha-robust mean-variance reinsurance-investment strategy. *Journal of Economic Dynamics and Control*, *70*, 101–123.

Li, D., Zeng, Y., & Yang, H. (2018). Robust optimal excess-of-loss reinsurance and investment strategy for an insurer in a model with jumps. *Scandinavian Actuarial Journal*, *2018*(2), 145–171.

Liang, Z., Chen, H., Zhu, J., Jiang, K., & Li, Y. (2018). Adversarial deep reinforcement learning in portfolio management. *arXiv preprint arXiv:1808.09940*.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lim, S. H., & Autef, A. (2019). Kernel-based reinforcement learning in robust markov decision processes. In *International conference on machine learning* (pp. 3973–3981).

Lim, S. H., Xu, H., & Mannor, S. (2013). Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, *26*.

Maccheroni, F., Marinacci, M., & Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, *74*(6), 1447–1498.

Maenhout, P. J. (2004). Robust portfolio rules and asset pricing. *The Review of Financial Studies*, *17*(4), 951–983.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, *51*(3), 247–257.

Milgrom, P., & Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, *70*(2), 583–601.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

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

Moody, J., Wu, L., Liao, Y., & Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, *17*(5-6), 441–470.

Nachum, O., Gu, S., Lee, H., & Levine, S. (2018). Data-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:1805.08296*.

Nevmyvaka, Y., Feng, Y., & Kearns, M. (2006). Reinforcement learning for optimized trade execution. *Proceedings of the 23rd international conference on Machine learning*, 673–680.

Nilim, A., & El Ghaoui, L. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, *53*(5), 780–798.

Olkin, I., & Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, *48*, 257–263.

Ormoneit, D., & Glynn, P. (2002). Kernel-based reinforcement learning in average-cost problems. *IEEE Transactions on Automatic Control*, *47*(10), 1624–1636.

Park, H., Sim, M. K., & Choi, D. G. (2020). An intelligent financial portfolio trading strategy using deep q-learning. *Expert Systems with Applications*, *158*, 113573.

Pendharkar, P. C., & Cusatis, P. (2018). Trading financial indices with reinforcement learning

agents. *Expert Systems with Applications*, *103*, 1–13.

Pirvu, T. A., & Zhang, H. (2014). Investment–consumption with regime-switching discount rates. *Mathematical Social Sciences*, *71*, 142–150.

Pliska, S. R., & Ye, J. (2007). Optimal life insurance purchase and consumption/investment under uncertain lifetime. *Journal of Banking & Finance*, *31*(5), 1307–1319.

Pulford, B. D. (2009). Short article: Is luck on my side? optimism, pessimism, and ambiguity aversion. *Quarterly Journal of Experimental Psychology*, *62*(6), 1079–1087.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Richard, S. F. (1975). Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model. *Journal of Financial Economics*, *2*(2), 187–203.

Rigter, M., Duckworth, P., Lacerda, B., & Hawes, N. (2022). Planning for risk-aversion and expected value in mdps. , *32*, 307–315.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. *Journal of risk*, *2*, 21–42.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., & Teh, Y. W. (2018). An analysis of categorical distributional reinforcement learning. *International Conference on Artificial Intelligence and Statistics*, 29–37.

Ruszczyński, A. (2010). Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, *125*, 235–261.

Schlosser, R. (2020). Risk-sensitive control of markov decision processes: A moment-based approach with target distributions. *Computers & Operations Research*, *123*, 104997.

Shen, Y., & Siu, T. K. (2013). Stochastic differential game, Esscher transform and general equilibrium under a Markovian regime-switching Lévy model. *Insurance: Mathematics and Economics*, *53*(3), 757–768.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... others (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning* (pp. 387–395).

Siu, T. K. (2011). Regime-switching risk: To price or not to price? *International Journal of Stochastic Analysis*.

Stella, X. Y., Lin, Y., & Yan, P. (1998). Optimization models for the first arrival target distribution function in discrete time. *Journal of mathematical analysis and applications*, *225*(1), 193–223.

Tamar, A., Chow, Y., Ghavamzadeh, M., & Mannor, S. (2016). Sequential decision making with coherent risk. *IEEE transactions on automatic control*, *62*(7), 3323–3338.

Tamar, A., Di Castro, D., & Mannor, S. (2012). Policy gradients with variance related risk criteria. , 387–396.

Tamar, A., Glassner, Y., & Mannor, S. (2015). Optimizing the cvar via sampling. , *29*(1).

Tamar, A., Mannor, S., & Xu, H. (2014). Scaling up robust mdps using function approximation. In *International conference on machine learning* (pp. 181–189).

Tang, Y. C., Zhang, J., & Salakhutdinov, R. (2020). Worst cases policy gradients. *Conference on Robot Learning*, 1078–1093.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Wang, M., & Ku, H. (2022). Risk-sensitive policies for portfolio management. *Expert Systems with Applications*, *198*, 116807.

Wang, P., & Li, Z. (2018). Robust optimal investment strategy for an AAM of DC pension plans with stochastic interest rate and stochastic volatility. *Insurance: Mathematics and Economics*, *80*, 67–83.

Wang, Y., & Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, *34*, 7193–7206.

Wang, Y., & Zou, S. (2022). Policy gradient method for robust reinforcement learning. In *International conference on machine learning* (pp. 23484–23526).

Wiesemann, W., Kuhn, D., & Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, *38*(1), 153–183.

Wu, J., & Li, H. (2020). Deep ensemble reinforcement learning with multiple deep deterministic policy gradient algorithm. *Mathematical Problems in Engineering*, *2020*, 1–12.

Xiong, Z., Liu, X.-Y., Zhong, S., Yang, H., & Walid, A. (2018). Practical deep reinforcement learning approach for stock trading. *arXiv preprint arXiv:1811.07522*.

Yarri, M. E. (1965). Uncertain lifetime, life insurance, and the theory of the consumer. *The Review of Economic Studies*, *32*(2), 90.

Yu, P., & Xu, H. (2015). Distributionally robust counterpart in markov decision processes. *IEEE Transactions on Automatic Control*, *61*(9), 2538–2543.

Zariphopoulou, T. (1992). Investment-consumption models with transaction fees and

Markov-chain parameters. *SIAM Journal on Control and Optimization*, *30*(3), 613–636.

Zhang, A., Sodhani, S., Khetarpal, K., & Pineau, J. (2020). Learning robust state abstractions for hidden-parameter block mdps. *arXiv preprint arXiv:2007.07206*.

Zhang, Y., & Desilva, C. W. (2014). Rsmdp-based robust q-learning for optimal path planning in a dynamic environment. *IAES International Journal of Robotics and Automation*, *3*(1), 1.

Zhou, X. Y., & Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic lq framework. *Applied Mathematics and Optimization*, *42*(1), 19–33.