

Some Continuous-Time Reinforcement Learning Problems in Finance

by

Yuling Max Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2026

© Yuling Max Chen 2026

Examining Committee Membership

The following members served on the Examining Committee for this thesis defense.

External Member: Cody Hyndman
Full Professor, Dept. of Mathematics and Statistics,
Concordia University

Internal-External Member: Alan Huang
Professor, School of Accounting and Finance,
University of Waterloo

Internal Members: Chengguo Weng
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

 Mario Ghossoub
Associate Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Supervisors: Bin Li
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

 David Saunders
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Reinforcement Learning (RL) is an emerging technique to solve for optimal policies of sequential decision-making problems, particularly when the parameters of the dynamic environment are unknown. Classical RL is framed under a discrete-time setting, often modeled on a Markov Decision Process. While recent RL literature has extended their problems to continuous-time, the lack of mathematical foundation limits their success to the empirical level. On the other hand, while classic stochastic control provides theoretically optimal solutions to continuous-time sequential decision-making problems, it relies on the model assumptions and the knowledge of true model parameters of the dynamic environment. Therefore, recent stochastic control literature has shown a growing interest in applying RL techniques to classical control problems and formulating traditional RL methodologies into rigorous stochastic control frameworks. By randomizing the control to a probability distribution, RL facilitates an informed exploration within the control space, allowing the decision-maker to seek for the optimal strategy without the full knowledge of the model of the dynamic environment. We refer to this RL-facilitated control randomization as an *exploratory extension* to the stochastic control problems, leading to an *exploratory control framework*.

In light of the wide applications of continuous-time RL and stochastic control, this thesis studies four classical stochastic control problems under the *exploratory control framework*, which is motivated by three mainstream finance problems — the mean-variance portfolio optimization problem (Chapters 2 and 3), the optimal investment-consumption problem (Chapter 4), and the optimal stopping problem with an application in exercising American-type options (Chapter 5).

Chapter 2 studies the mean-variance portfolio optimization problem in a regime-switching market. By adopting the Lagrangian dual of the mean-variance objective, we transform the mean-variance problem to a time-consistent control problem and solve for a precommitted Gaussian investment policy that is globally optimal at the initial time. Whereas, Chapter 3 revisits the mean-variance portfolio optimization problem in a market with jumps and directly optimizes the original mean-variance objective, where time inconsistency arises. As a result, we solve for an equilibrium Gaussian investment policy that is locally optimal in a game-theoretical equilibrium sense. Chapter 4 addresses Merton’s optimal investment-consumption problem with non-exponential discounting, which is a time-inconsistent dual-control problem. We solve for an equilibrium investment-consumption policy, as a joint distribution over the investment and consumption space, and further show that it can be expressed as the product of a Gaussian investment policy and a Gamma consumption policy. Finally, Chapter 5 tackles the optimal stopping problem and accounts for a gen-

eral (not necessarily exponential) discounting function. With the binary character of the control space, stop and continue, we solve for an equilibrium Bernoulli stopping policy, which represents the probability of stopping. We also examine the effect of exponential and non-exponential discounting on the optimal stopping policies and the stopping regions.

For each problem, we solve for the optimal policy and the corresponding value functions, analytically or semi-analytically. We provide verification theorems and policy improvement (or policy iteration) theorems, as well as supplementary lemmas and propositions, which serve as the theoretical foundation of the subsequent analysis. Beyond these, we also design an RL algorithm for each problem. We derive effective loss functions that apply the martingale properties of the theoretical solutions. With proper RL model configurations, our algorithms have empirically shown their capability of learning the optimal policies in our numerical studies with simulated and real market data.

Acknowledgements

I extend my heartfelt gratitude to my supervisors, Dr. Bin Li and Dr. David Saunders, for their invaluable guidance throughout my research journey and their thoughtful reviews of all my projects, including this thesis. Their expertise and support have been instrumental in shaping my academic work. Their education and cultivation have pruned my personality and capability, guiding me to become a meticulous, diligent and resilient man. I am equally thankful to the department for fostering a collaborative and resourceful academic environment, providing me with the necessary tools, assistance, and opportunities to pursue research seamlessly and effectively. I also deeply appreciate the donors of Ontario Grad Scholarship and James Downey Grad Scholarship, for their generosity and recognition that supported my research and helped me overcome the difficult times, both financially and emotionally. I also express my gratefulness to my family, including but not limited to, my mother, my wife, my son and my cat, for their persistent support and long-lasting accompaniment during this challenging and sometimes arduous journey of pursuing a doctorate degree.

Without any one above, it will not be possible for me to reach this point.

Dedication

This is dedicated to my mother, Wanyun Zhao, my wife, Meiqi Shu, my son, Wencheng Nolyn Chen and my cat, Ruyee Chen.

Table of Contents

Examining Committee Membership	ii
Author's Declaration	iii
Abstract	iv
Acknowledgements	vi
Dedication	vii
List of Figures	xii
List of Tables	xiv
1 Introduction	1
2 Exploratory Mean-Variance Portfolio Optimization with Regime-Switching Market Dynamics	7
2.1 Introduction	7
2.2 Classical Mean-Variance with Regime-Switching Problem	10
2.3 Exploratory Mean-Variance with Regime-Switching Problem	13
2.4 RL Algorithm	19
2.4.1 Temporal Difference (TD) Learning	20

2.4.2	Orthogonality Condition (OC) Learning	22
2.4.3	Updating Scheme and Training Algorithm	24
2.4.4	Training on Real Market Data	27
2.5	Numerical Results	28
2.5.1	Simulation Study: Comparing TD and OC Learning	28
2.5.2	Performance on Real Market Data: Comparing EMVRS and EMV	33
2.6	Conclusion	36
3	Exploratory Mean-Variance with Jumps: An Equilibrium Approach	39
3.1	Introduction	39
3.2	Problem Formulation and Preliminaries	42
3.3	Exploratory Mean-Variance with Jumps Problem	45
3.4	RL Algorithm	49
3.5	Numerical Results	52
3.5.1	Analysis On Simulated Market Data	53
3.5.2	Analysis on Real Market Data	56
3.6	Conclusion	60
4	Exploratory Investment-Consumption Problem with Non-Exponential Discounting	62
4.1	Introduction	62
4.2	Classical Optimal Investment-Consumption Problem	66
4.3	Exploratory Investment-Consumption Problem	70
4.4	Theoretical Results	73
4.5	RL Algorithm	80
4.5.1	Policy Evaluation Approach	82
4.5.2	q-Learning Approach	83
4.6	Numerical Results	86
4.7	Conclusion	89

5	Equilibrium Exploratory Stopping Problem with Non-Exponential Discounting	91
5.1	Introduction	91
5.2	Optimal Stopping with Precommitted Strategy	94
5.2.1	Classical Optimal Stopping	95
5.2.2	Exploratory Extension to Classical Optimal Stopping	97
5.2.3	Time Inconsistent Issue in Classical Optimal Stopping	99
5.3	Equilibrium Exploratory Optimal Stopping with Non-Exponential Discounting	100
5.3.1	Equilibrium Policy and Verification Theorem	101
5.3.2	Comparison Between Exponential and Non-Exponential Discounting	106
5.4	RL and Finite Difference Algorithms	108
5.5	Numerical Results	115
5.6	Conclusion	120
6	Conclusion and Future Works	123
	References	129
A	APPENDICES	138
	APPENDICES	138
A.1	Proofs and Derivations of Chapter 2	138
A.1.1	Derivation of (2.14)	138
A.1.2	Derivation of (2.20)	139
A.1.3	Proof of Theorem 2.3.1	140
A.1.4	Proof of Corollary 2.3.1	145
A.1.5	Proof of Theorem 2.3.2	146
A.2	Proofs of Chapter 3	147
A.2.1	Proof of Lemma 3.2.0.1	147

A.2.2	Proof of Theorem 3.3.1	148
A.2.3	Proof of Theorem 3.3.2	150
A.2.4	Proof of Theorem 3.4.1	153
A.3	Proofs of Chapter 4	155
A.3.1	Proof of Proposition 4.2.1	155
A.3.2	Proof of Theorem 4.4.1	155
A.3.3	Proof of Lemma 4.4.0.1	158
A.3.4	Proof of Theorem 4.4.2	160
A.3.5	Proof of Corollary 4.4.1	163
A.3.6	Proof of Theorem 4.4.3	165
A.3.7	Proof of Theorem 4.5.1	166
A.3.8	Proof of Theorem 4.5.2	168
A.4	Proofs of Chapter 5	170
A.4.1	Derivation of (5.26) and (5.27)	170
A.4.2	Proof of Theorem 5.3.1	172
A.4.3	Proof of Proposition 5.3.1	174
A.4.4	Proof of Proposition (5.3.2)	177
A.4.5	Proof of Theorem 5.4.1	179
A.4.6	Proof of Corollary 5.4.1	179

List of Figures

1.1	RL Workflow	1
2.1	Parameter convergence of Temporal Difference (TD) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.8, -0.3)$. Learning errors of each parameter are (400%, 400%, 112.11%, 77.92%).	30
2.2	Parameter convergence of Orthogonality Condition (OC) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.8, -0.3)$. Learning errors of each parameter are (1.75%, 0.5%, 0.79%, 0.56%).	31
2.3	Parameter convergence using Orthogonality Condition (OC) loss. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.2, 0.2)$. Learning errors of each parameter are (0.95%, 1.25%, 0.14%, 0.18%).	31
2.4	Parameter convergence of Orthogonality Condition (OC) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.5, -0.5)$. Learning errors of each parameter are (0.25%, 5.07%, 4.45%, 0.34%).	33
2.5	Regime-identified financial time series for training (left: S&P500, right: 3mTbill). State 0 stands for bullish and state 1 stands for bearish.	34
3.1	Convergence of μ (left), σ (middle) and δ (right).	54
3.2	S&P500 market index. The period between the dotted vertical lines is the year 2018 and the period between the dashed vertical lines is the year 2022. The market was significantly shocked due to the crash of the cryptocurrency bubble in 2018, as well as the post-COVID historical global interest rate hike and the escalation in the Russo-Ukrainian war in 2022.	60
4.1	Parameter convergence using non-exponential discounting, μ (left) and σ (right).	88

4.2	Parameter convergence using exponential discounting, μ (left) and σ (right).	88
5.1	Free boundaries from different initial times $t_0 = 0$ vs $t_0 = 0.2$ for exponential (left) and non-exponential (right) discounting.	100
5.2	Discounting functions (left) and their decision boundaries (right).	117
5.3	The decision boundaries learned via the RL algorithm are characterized by the edges between the “exercise” (green) and “do not exercise” (red) points, under exponential discounting (left) and non-exponential discounting (right) settings. The free boundaries (blue curves) are numerically solved via the finite difference method and are considered as the benchmark.	118
5.4	Empirical densities (left) and CDFs (right) of exit times for exponential and non-exponential discounting.	119
5.5	Empirical densities of average exit times for exponential and non-exponential discounting.	120

List of Tables

2.1	Configuration of the pilot simulation.	29
2.2	Configuration of the more comprehensive simulation.	32
2.3	Configuration of the real data study.	35
2.4	Performance on the 24 10-year windows. Mean and Volatility stand for the annualized mean and volatility of portfolio returns on the 24 10-year windows. AC is action constraint, SS is short selling, and SR is Sharpe Ratio. The target annual portfolio return is given by 3.422%.	37
3.1	Parameters for training and evaluation on simulated data.	54
3.2	Investment performance on simulated market data. The Monte Carlo standard errors of the realized means, realized volatilities and $J_{empirical}$ are reported in parentheses.	56
3.3	Investment performance on training periods. The model was trained with $(\lambda, \gamma) = (5, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.01, 5)$. The Monte Carlo standard errors of the portfolio value means and volatilities are reported in parentheses.	57
3.5	Investment performance in 2018. The model was retrained with $(\lambda, \gamma) = (1, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.01, 5)$. SR denotes the Sharpe Ratio. The Monte Carlo standard errors of the portfolio value mean and volatility are reported in parentheses.	58
3.4	Investment performance on evaluation periods. The model was trained with $(\lambda, \gamma) = (5, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.1, 5)$. SR denotes the Sharpe Ratio, and the subscript MLE denotes the investment performance calculated using MLEs of Merton's jump-diffusion model. The Monte Carlo standard errors of the portfolio value means and volatilities are reported in parentheses.	59

5.1	RL model configuration. Total number of trainable parameters: 481,171.	117
5.2	Discounted option payoffs.	119

Chapter 1

Introduction

Reinforcement Learning (RL) is a branch of machine learning that focuses on optimizing sequential decision-making under uncertainty, where a decision-maker interacts with an environment to maximize cumulative rewards. As in Figure 1.1, at each time t , an *agent*, i.e., the decision-maker, finds herself at *state* S_t in the dynamic environment. By selecting an *action* a_t based on the *policy* π , the *agent* receives the *reward* R_t and transitions to the next *state* S_{t+1} . The *agent's* objective is to maximize the cumulative *reward*, which is often discounted over time. As introduced by (Sutton & Barto, 2018), a RL agent maintains an *exploration-exploitation trade-off*, balancing between taking the known most-rewarding action (exploitation) and trying some unknown but potentially more rewarding actions (exploration).

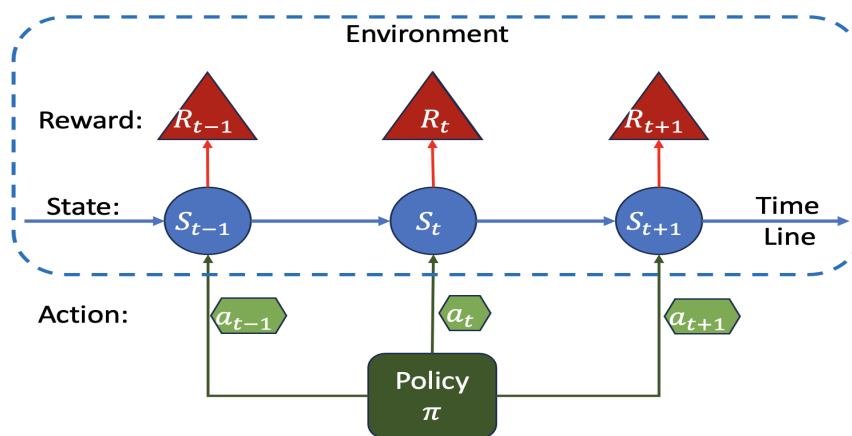


Figure 1.1: RL Workflow

While traditional RL operates in discrete-time settings, continuous-time RL has gained attention for applications involving high-frequency decision-making and stochastic processes. Continuous-time RL leverages mathematical tools such as dynamic programming, Hamilton-Jacobi-Bellman (HJB) equations, and functional approximations to address problems where decisions occur in a continuous-time framework (Doya, 2000). In turn, the stochastic control literature has shown growing interest in applying continuous-time RL techniques to the classical control problems (H. Wang et al., 2020; Lesmana & Pun, 2021), as well as formulating existing RL methodologies into more rigorous stochastic control frameworks (Jia & Zhou, 2022a, 2022b, 2023). The key idea of applying RL techniques to stochastic control problems is to introduce stochasticity into the control, randomizing it from a deterministic function (that maps time and state to the control space) to a probability distribution over the control space conditioned on time and state. This idea was first pointed out by Gullapalli (1990), who studied the stochastic RL in discrete time, and then advanced by H. Wang et al. (2020), who formulated the *exploratory control framework* in continuous time. We hereafter refer to this control randomization as an *exploratory extension* to the classical control problem, and refer to the control problem after extension as an *exploratory control problem*.

The reason behind such exploratory extension is two-folded. First and foremost, it is often unrealistic to have full access to the model that drives the dynamic environment. For example, an investor will not be able to observe the true mean and volatility of a stock's return in reality, if the stock price dynamic is driven by a geometric Brownian Motion. Yet, most stochastic control solutions are optimal provided that the parameters are known. So, a RL policy is a randomized exploratory control, often defined as a probability distribution centered around the classical control solution, with some variance that determines the level of exploration within the control space. In other words, RL enables us to effectively explore within the control space, and find the optimal policy through iterative interactions with the dynamic environment, without the knowledge of the true model parameters of the dynamic environment. Second, classical approaches to stochastic control problems involve characterization of the value function from the HJB equation, a Partial Differential Equation (PDE) that can be solved by numerical methods (Press et al., 2007). Yet, when the model of the dynamic environment is complex and the *state* is high-dimensional, the PDE easily becomes intractable and hence cannot be handled by classical numerical solvers. RL maintains the compatibility to deep learning techniques, known as *deep RL*, which parameterizes the value function and the policy with neural networks and trains them iteratively via the Actor-Critic Algorithm (Sutton & Barto, 2018); See more examples of deep RL from (Arulkumaran et al., 2017). More recent works (Jia & Zhou, 2022a, 2022b, 2023) further fulfilled the theoretical foundation of RL, proposing effective loss functions

that are derived from the martingale properties of the stochastic control solutions.

On top of that, the literature also considers entropy regularization, incorporating the entropy of an exploratory control (as a probability distribution) to the objective function, which is motivated by two additional advantages of the entropy regularizer. First, maximizing an objective with entropy regularization (or equivalently minimizing an objective with a negative entropy term) enforces the exploratory control to avoid collapsing too fast to a single deterministic control. Thus, it encourages policy exploration within the control space and balances the exploitation-exploration trade-off. Second, the entropy term effectively regularizes the induced policies to the Boltzmann family of distributions, yielding a Gaussian policy in the Linear-Quadratic setting and a Gamma policy in the logarithmic Hamiltonian setting for instance. The resulting policies often possess well-known statistical properties and are easier to learn with a RL model. So, we adopt entropy regularization throughout this thesis.

Additionally, finance is a prominent application domain for RL, where the dynamic and stochastic nature of markets aligns well with RL’s problem formulation. An investor can be regarded as the aforementioned *agent* and the market corresponds to the *state* of the dynamic environment. The investor’s investment strategy is a *policy* defined as a probability distribution, from which an investment action is executed (i.e., sampled from this distribution). Then, the goal of this game is to find the investment strategy that optimizes an investment objective, often defined as the discounted cumulative reward. RL techniques have been applied to various finance problems, such as portfolio optimization (H. Wang & Zhou, 2020; R. Jiang et al., 2022; Dai et al., 2023; B. Wu & Li, 2024; Chen et al., 2025b) and pricing and hedging of American-type options (Becker et al., 2020; Dai et al., 2024).

In the subsequent Chapters, we apply RL techniques to three well-studied finance problems, the mean-variance portfolio optimization problem (Chapters 2 and 3), the optimal investment-consumption problem (Chapter 4) and the optimal stopping problem with an application of exercising American options (Chapter 5). In each Chapter, we review the literature and their classical control solutions, based on which we derive the exploratory version of these problems. We solve for the optimal policy, either analytically or semi-analytically depending on the solvability of the exploratory extended HJB equations. Besides, we also develop RL algorithms for each problem, propose effective loss functions based on the martingale properties of the theoretical solutions, and empirically examine our algorithms with numerical studies. It is worth mentioning that the RL models in Chapters 2, 3 and 4 extensively leverage the analytical solutions and are parameterized by the model parameters of the dynamic environment, which are the coefficients of diffu-

sion models. Hence, a crucial evaluation criterion for these RL models is whether the RL models’ parameters converge to the corresponding “grounding true” values of the dynamic environment’s model parameters, and of course, they successfully did. That indicates our RL models’ capability to unravel the underlying model dynamics of the environment, via sequential interactions with the environment and without any parameter estimation involved.

More specifically, Chapter 2 studies the mean-variance portfolio optimization problem in a regime-switching market. Following (Zhou & Yin, 2003) who solved this problem under the classical control setting, we extend it to the exploratory control framework, termed as the Exploratory Mean Variance with Regime Switching (EMVRS) problem. With entropy regularization, the investment policy transforms from the classical solution (as a deterministic function) to a Gaussian distribution centered around the classical solution. The optimal investment policy is able to make proper adjustments contingent to the market regimes, such as investing more prudently in a regime with higher market volatility. We also solve the corresponding value function semi-analytically, which depends on the solutions to a system of four Ordinary Differential Equations (ODEs). Meanwhile, this Chapter can also be viewed as an extension of (H. Wang & Zhou, 2020), who first studied the Exploratory Mean-Variance (EMV) problem, to a more complicated market with finite Markovian regimes. Furthermore, this Chapter considers and focuses on the Lagrangian dual of the mean-variance problem, as similarly addressed in (Zhou & Yin, 2003) and (H. Wang & Zhou, 2020), which bypasses the time-inconsistency arising from the original mean-variance objective function. As a result, our optimal investment policy is a precommitted strategy. In a numerical study, we compare the investment performance of our EMVRS investor with an EMV investor that does not account for the market regimes, and the EMVRS investor achieves superior performance on real market data. Finally, the main results of this Chapter have been published at the International Journal of Theoretical and Applied Finance; see (Chen et al., 2025b).

Chapter 3 revisits the mean-variance portfolio optimization problem, considering a market with jumps. Past works such as (Øksendal & Sulem, 2005) have studied the mean-variance problem with jump-diffusion controlled dynamics in the classical, non-exploratory control formulation, while Bender and Thuan (2023) extended it to the exploratory control framework. Gao et al. (2024) further introduced the exploratory formulation of a general stochastic control problem with jump-diffusion controlled dynamics. In contrast to (Øksendal & Sulem, 2005; Bender & Thuan, 2023) and Chapter 2 that consider the Lagrangian dual of the mean-variance problem, this Chapter directly addresses the time-inconsistency arising from the variance term in the the original mean-variance objective function. Following an equilibrium approach as well-noted by (Björk et al., 2021), we solve

for an equilibrium investment policy, which is optimal in a game-theoretical equilibrium sense and is different from a precommitted strategy as in (Øksendal & Sulem, 2005; Bender & Thuan, 2023; Gao et al., 2024). Under the exploratory framework with entropy regularization, an equilibrium investment policy is a Gaussian distribution, that is aware of the market jumps — making more conservative investment decisions when the market jump has high intensity and magnitude. It is noteworthy that this Chapter is a complementary work to (Dai et al., 2023), who also studied an equilibrium mean-variance portfolio optimization strategy and introduced an additional exogenous state variable to model the unobservable economic conditions. In a numerical study, the proposed equilibrium investment policy shows strong performance on real market data and is robust against the shift of market trends. Finally, the major results of this Chapter have been accepted by a commemorative publication of Dr. Rudi Zagst; see (Chen et al., 2025c).

Chapter 4 addresses Merton’s optimal investment-consumption problem with non-exponential discounting. We remark that this is a dual-control problem, which is essentially different from the past works in the exploratory control framework (including Chapter 2 and 3) that solve for a single control such as investment. Notably, even an investment across various assets is still considered as a single control with multiple dimensions. Furthermore, the presence of non-exponential discounting adds an additional layer of complexity as it introduces time-inconsistency, which models people’s time-varying preferences. Adopting the log utility objective function from (Björk et al., 2021) and following their equilibrium approach, we extend the problem to the exploratory control framework and solve for an equilibrium investment-consumption policy. With entropy regularization, our equilibrium investment-consumption policy, as a joint distribution over the investment and consumption space, can be written as a product of a Gaussian investment policy and a Gamma consumption policy. It is noteworthy that the investment, defined as the portfolio weight in a risky asset, can take positive and negative values respectively corresponding to leverage-buying and short-selling, whereas consumption has to be nonnegative. Meanwhile, the domain of a Gaussian distribution is $\mathbb{R}$ and the domain of a Gamma distribution is $\mathbb{R}_+$, which respectively align with the domains of investment and consumption in practice. Hence, our equilibrium investment-consumption policy is practically meaningful. Furthermore, this Chapter also investigates the change in behaviors of investors using exponential and non-exponential discounting. Following the assumptions on non-exponential discounting functions from (Ebert et al., 2020), we find out that a non-exponential discounting investor will behave more conservatively in both consumption and investment under the exploratory framework, whereas she will only show her prudence in consumption under the classical, non-exploratory framework. Finally, the main result of this Chapter has been submitted to SIAM Journal of Financial Mathematics and is currently under review.

Chapter 5 tackles an optimal stopping problem with non-exponential discounting in finite horizon, with an application in exercising an American put option. This Chapter extends (Dai et al., 2024), who studied an exploratory stopping problem under a time-consistent setting, with the incorporation of a general, not necessarily exponential, discounting function. With the possibly non-exponential discounting, we consider a time-inconsistent setting and solve for an equilibrium stopping policy following (Björk et al., 2021). While the time-inconsistent optimal stopping problem involves characterization of value functions from systems of free boundary PDEs, we adopt the penalty approximation method, as also applied in (Dai et al., 2024), and transform the free boundary PDEs into penalized PDEs that can be viewed as the extended HJB equations of a standard time-inconsistent control problem. Then, we extend the problem to the exploratory control framework and, with entropy regularization, solve for a Bernoulli equilibrium stopping policy, which represents the probability of stopping. Furthermore, we compare the effect of exponential and non-exponential discounting on the solutions and find out two interesting results: i) in the exploratory framework, non-exponential discounting yields lower probability of stopping; and ii) in the classical, non-exploratory setting, non-exponential discounting yields smaller stopping regions, both under the assumption that the means of the rates of the non-exponential and exponential discounting functions are equal. Both findings, along with the numerical examples, depict a more patient non-exponential discounting agent, willing to wait for a potentially better future payoff. In addition to the RL algorithm, we also develop a finite difference algorithm to solve for the value function from the extended HJB equations, which relies on the true model parameters of the dynamic environment. In a numerical study based on an American put option, we apply the numerical algorithm to solve for the free boundary, which is the early exercise price at each time before maturity, and set it as the benchmark for our RL algorithm. Our RL model successfully calibrates this free boundary via repeatedly implementing the Bernoulli equilibrium stopping policy over batches of simulated stock price trajectories and learning from its historical practices. Finally, the main result of this Chapter has been submitted to Quantitative Finance of Taylor & Francis Group and is currently under review.

Standing on the shoulders of the giants, we acknowledge that this work will not be completed without the dedicated effort of past researchers in this field. With some humble thoughts on the RL application in finance, we conclude this thesis in Chapter 6 with some high-level review and prospect on future works.

Chapter 2

Exploratory Mean-Variance Portfolio Optimization with Regime-Switching Market Dynamics

2.1 Introduction

Mean-Variance (MV) portfolio optimization has been widely studied since being introduced in (Markowitz, 1952). In the continuous-time setting, an investor with the classical MV objective aims to simultaneously maximize their expected portfolio value and minimize the portfolio volatility at the end of the investment horizon. This has been studied as a stochastic linear-quadratic problem by Zhou and Li (2000), followed by Chiu and Li (2006) who solved for the MV efficient frontier, and Xie et al. (2008) who derived the optimal investment policy under an incomplete market setting; see (Zhang et al., 2018; Kalayci et al., 2019) for a broader overview of the the past work in the MV literature.

Amongst many variants of the MV problem, this chapter focuses on the regime-switching setting. Past works such as (Maheu & McCurdy, 2000; Ang & Bekaert, 2002, 2004) have demonstrated that incorporating a regime-switching model better fits real market data and that ignoring market regimes in investment has a cost. Yin and Zhou (2004) considered a discrete-time MV problem with regimes modeled on an aggregated process and asymptotically derived the optimal portfolio selection strategy from the optimal solution of its continuous-time counterpart. H. Wu and Li (2011) and H. Wu et al. (2014) solved regime-switching MV portfolio allocation problems with an uncertain investment horizon,

where the time to exit the market is modeled by a conditional distribution given the market regime.

Recent research in stochastic control (H. Wang & Zhou, 2020; R. Jiang et al., 2022; Jia & Zhou, 2022b, 2023; Dai et al., 2023; Denkert et al., 2024; B. Wu & Li, 2024) has shown the advantage of adopting Reinforcement Learning (RL) techniques into classical stochastic control problems. The basic strategy is to employ the Stochastic RL Algorithm introduced in (Gullapalli, 1990), that replaces the optimal solution to the classical stochastic control problem with a probability distribution. By introducing stochasticity to the optimal control solution, we enable exploration within the control space, while maintaining exploitation towards optimality. This allows the induced optimal policy distribution to be more robust against the randomness of the dynamic environment. H. Wang et al. (2020) introduced the “exploratory extension” of the portfolio value dynamic and solved for the corresponding exploratory optimal solution to the linear-quadratic problem, where they proved the asymptotic equivalence between the classical solution and the exploratory solution. Later, H. Wang and Zhou (2020) introduced the *Exploratory Mean-Variance* (EMV) problem and solved for the optimal investment strategy, which is a probability distribution over the control space rather than a deterministic function of the control.

In this chapter, we study the RL-facilitated EMV problem by (H. Wang & Zhou, 2020) in the context of a regime-switching process, named as the *Exploratory Mean Variance with Regime Switching* (EMVRS) problem. The continuous-time MV portfolio optimization problem with regime-switching but without RL has previously been studied by (Zhou & Yin, 2003), which we hereafter refer to as the *Mean Variance with Regime Switching* (MVRS) problem. On top of that, we design a value-based RL algorithm that reparameterizes the analytical solution to the EMVRS problem as the value function. In contrast to defining the value function as a randomly initialized neural network or polynomial approximator such as in (Z. Jiang et al., 2017; Wen et al., 2021; B. Wu & Li, 2024), we configure a RL model whose parameters all have practical meanings. Finally, we applied the martingale property of the deduced optimal EMVRS value function, specifically the *Orthogonality Condition* studied by (Jia & Zhou, 2022a), to the updating scheme of the parameters in the RL algorithm.

We make contributions to the literature by generalizing the EMV problem (H. Wang & Zhou, 2020) to the regime-switching market setting and extending the analytical solution of the MVRS problem (Zhou & Yin, 2003) with the RL policy exploration. In contrast to the Temporal Difference (TD) learning in (H. Wang & Zhou, 2020), our Orthogonality Condition (OC) learning algorithm (Algorithm 1) achieves convergence to the true parameters in our simulation study; see Figure 2.1 and 2.2 for comparison. On the real market data, the EMVRS model trained with OC learning reaches the highest Sharpe Ratio com-

pared to the other models that trained with TD learning or that only considers a single regime; see Table 2.4 for back-testing investment performance of EMVRS and EMV on the real market data.

Furthermore, B. Wu and Li (2024) also studied the RL-facilitated MV portfolio selection problem with a regime switching setting. B. Wu and Li (2024) formulated the problem as a Partially Observable Markov Decision Process with two unobservable market regimes, estimating the regimes using the Wonham filter (Eq. (10) of B. Wu and Li (2024)). There are three major differences between our work and (B. Wu & Li, 2024). Firstly, B. Wu and Li (2024) assumed that the volatility is regime-independent on a two-regime market, while we allow for regime-dependent volatility and a multiple-regime market. Secondly, the RL algorithm in (B. Wu & Li, 2024) requires the market parameters¹ to be either given or estimated from the data, and selects the Martingale Loss, first introduced in (Jia & Zhou, 2022a), to update the RL model parameters. We notice from (Jia & Zhou, 2022a) that the Martingale Loss requires the knowledge of the true market parameters (which is unavailable in the real market), whereas the Orthogonality Condition Loss does not. According to the Mean-Blur problem² arguments in (Luenberger, 2013), it is possible that replacing the market parameters with their estimates can cause the RL model to update towards an erroneous target. Therefore, by adopting the Orthogonality Condition Loss, our RL algorithm does not require the knowledge of the market parameters, instead, it learns the market parameters through training. Thirdly, by parameterizing the RL model with the market parameters, we show in our simulation study that the parameters of our proposed RL model can converge to their corresponding true values, with randomly chosen starting point; see Figure 2.3.

The remainder of the chapter is structured as follows. In Section 2.2, we present background material and the problem setup. Then, we introduce the EMVRS problem, provide an analytical solution, and establish a Policy Improvement Theorem in Section 2.3. In Section 2.4, we present a reparameterization of the EMVRS problem, followed by RL algorithms to update the market parameters. We present numerical results on simulated and real market data in Section 2.5. Section 2.6 concludes and provides directions for future research.

¹The parameters of the market dynamics and regime dynamics correspond to $\tilde{\sigma}, \tilde{\mu}_i, \tilde{\lambda}_i$ for $i = 1, 2$ in Algorithm 1 of (B. Wu & Li, 2024).

²The Mean-Blur problem in quantitative finance refers to the extreme difficulty in estimating an asset's true expected return (mean) using historical data, as it is nearly impossible to calculate it within a workable level of accuracy.

2.2 Classical Mean-Variance with Regime-Switching Problem

In this section, we review the MVRS problem by (Zhou & Yin, 2003), which is a Lagrangian dual formulation of Markowitz's Mean-Variance (MV) Problem with a regime-switching market. Zhou and Yin (2003) solved this problem analytically without applying RL techniques. The Lagrangian dual formulation of the MV problem was also studied in (H. Wang & Zhou, 2020) with RL application, although they did not consider regime-switching. We start from the MVRS problem formulation.

Consider an investor who manages a portfolio with an investment horizon $T > 0$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space supporting a one-dimensional Brownian Motion $\{W_t\}_{t \in [0, T]}$ and a continuous-time stationary Markov chain $\{\alpha_t\}_{t \in [0, T]}$, which takes values in the finite state space $\{1, \dots, l\}$, where l represents the total number of the market regimes. The Brownian motion and the Markov chain are mutually independent. The Markov chain has the transition matrix $P(t) = (p_{ij}(t))_{i=1, \dots, l}^{j=1, \dots, l}$, where $p_{ij}(t) := \mathbb{P}(\alpha_t = j | \alpha_0 = i)$, and a generator $Q = (q_{ij})_{i=1, \dots, l}^{j=1, \dots, l}$ defined as $q_{ij} = \lim_{t \rightarrow 0+} t^{-1}(p_{ij}(t) - \delta_{ij})$ with δ_{ij} the Kronecker delta. It is so-defined such that the transition matrix is the matrix exponential of the generator, i.e., $P(t) = e^{tQ}$. Let $\{\mathcal{F}_t\}_{t \in [0, T]}$ be the filtration generated by $\{W_t\}_{t \in [0, T]}$ and $\{\alpha_t\}_{t \in [0, T]}$.

We assume the market consists of one risky asset, the stock $\{S_t\}_{t \in [0, T]}$, and one risk-free asset, the bond $\{B_t\}_{t \in [0, T]}$. The dynamics of the stock and the bond are driven by two stochastic processes

$$dS_t = S_t \{\mu(t, \alpha_t)dt + \sigma(t, \alpha_t)dW_t\}, \text{ with } S_0 > 0 \quad (2.1)$$

$$dB_t = r(t, \alpha_t)B_t dt, \text{ with } B_0 > 0, \quad (2.2)$$

where $\mu(t, i) \in \mathbb{R}$ and $\sigma(t, i) \in \mathbb{R}_+$ are respectively the mean and volatility of the stock return and $r(t, i) \in \mathbb{R}_+$ is the risk-free interest rate, at time $t \in [0, T]$ in market regime $i \in \{1, \dots, l\}$. We assume that $\mu(t, i) \neq \mu(t, j)$ and $\sigma(t, i) \neq \sigma(t, j)$, for any $t \in [0, T]$ and $i \neq j$.

It is obvious that $\{S_t\}_{t \in [0, T]}$ is adapted to $\{\mathcal{F}_t\}_{t \in [0, T]}$, the filtration generated by α and W , i.e., $\mathcal{F}_t^S \subset \mathcal{F}_t$, where $\{\mathcal{F}_t^S\}_{t \in [0, T]}$ is filtration generated by S . Conversely, we have $\mathcal{F}_t \subset \mathcal{F}_t^S$, meaning that the sample paths of α and W can be inferred from the historical prices of the stock. In fact, (2.1) shows that the market regime α can be deduced from the quadratic variation of the stock price as

$$d[S]_t = S_t^2 \sigma(t, \alpha_t)^2 dt.$$

Once the sample paths of α are known, one can further infer the drift process $\{\mu(t, \alpha_t)\}_{t \in [0, T]}$ and the Brownian motion $\{W_t\}_{t \in [0, T]}$.

At each time $t \in [0, T]$, denote X_t as the investor's portfolio value. The investor reallocates their portfolio by investing the amount u_t (the *control*) in the stock and $X_t - u_t$ in the bond. Under the self-financing assumption, the *portfolio value process* can be derived as

$$\begin{aligned} dX_t^u &= [r(t, \alpha_t)X_t^u + [\mu(t, \alpha_t) - r(t, \alpha_t)]u_t] dt + \sigma(t, \alpha_t)u_t dW_t \\ &= [r(t, \alpha_t)X_t^u + \rho(t, \alpha_t)\sigma(t, \alpha_t)u_t] dt + \sigma(t, \alpha_t)u_t dW_t, \end{aligned} \quad (2.3)$$

given the initial portfolio value $X_0 = x_0 > 0$ and the initial regime $\alpha_0 = i_0 \in \{1, \dots, l\}$. Here, $\rho(t, \alpha_t) := \sigma^{-1}(t, \alpha_t)(\mu(t, \alpha_t) - r(t, \alpha_t))$ is the Sharpe ratio, and $\{X_t^u\}_{t \in [0, T]}$ with a superscript u represents the portfolio value process that follows the *policy* $u := \{u_t\}_{t \in [0, T]}$. We next define the admissible control and state the classical Markowitz MV Problem.

Definition 2.2.1 (Admissible Control) *We say that a control u_t is admissible, denoted as $u_t \in \mathcal{A}$, if it satisfies the following conditions:*

- (i) $u_t : [0, T] \mapsto \mathbb{R} \in L^2_{\mathcal{F}}(0, T)$, i.e., u is a $\mathbb{R}$ -valued, $\{\mathcal{F}_t\}_{t \in [0, T]}$ -adapted function such that $\mathbb{E} \left[\int_0^T |u_t|^2 dt \right] < +\infty$;
- (ii) the SDE (2.3) has a unique solution X^u corresponding to u .

We denote a policy $u := \{u_t\}_{t \in [0, T]} \in \mathcal{A}$ if $u_t \in \mathcal{A}$, $\forall t \in [0, T]$.

Problem 2.2.1 (Classical Markowitz MV Problem)

$$\min_{u \in \mathcal{A}} \text{Var}(X^u(T)) \text{ subject to } \mathbb{E}(X^u(T)) = z, \quad (2.4)$$

where $z > 0$ is a prespecified target for the expected terminal wealth, i.e., the expected wealth to be achieved at the end of the investment horizon.

Noticing that the classical MV problem (Problem 2.2.1) is a constrained optimization problem, we consider its Lagrangian dual. Moreover, under the regime-switching market setting, we specify that the expectation in the original MV problem is actually conditioned on the initial states, including the investor's initial portfolio value $x_0 > 0$ and the initial market regime $i_0 \in \{1, \dots, l\}$.

Following the standard Dynamic Programming Principle (DPP) arguments, Zhou and Yin (2003) proposed and solved the MVRS problem (Problem 2.2.2), which is the Lagrangian dual of Problem 2.2.1. This modification reforms a time-inconsistent stochastic control problem into a time-consistent one, which makes it possible to obtain a precommitted solution.

Problem 2.2.2 (The MVRS Problem)

$$\min_{u \in \mathcal{A}} \left\{ J(u, x_0, i_0; \lambda) := \mathbb{E}[(X_T^u + \lambda - z)^2 | X_0^u = x_0, \alpha_0 = i_0] - \lambda^2 \right\}, \quad (2.5)$$

where $\lambda > 0$ is the Lagrangian multiplier, $z > 0$ is a prespecified target for the expected terminal wealth and $J(u, x_0, i_0; \lambda)$ is the objective function.

Define the *value function* as the infimum of the objective function over all admissible controls $u \in \mathcal{A}$. For $0 \leq t < s \leq T$, $X_t = x$, $\alpha_t = i$, and $u \in \mathcal{A}$, the *value function* is given by

$$V(t, x, i) = \inf_{u \in \mathcal{A}} \mathbb{E}[(X_T^u + \lambda - z)^2 | X_t = x, \alpha_t = i] - \lambda^2 \quad (2.6)$$

$$= \inf_{u \in \mathcal{A}} \mathbb{E}[V(s, X_s^u, \alpha_s) | X_t = x, \alpha_t = i]. \quad (2.7)$$

Following DPP arguments, the value function can be deduced to satisfy the Hamilton-Jacobi-Bellman (HJB) Equation

$$\begin{aligned} & v_t(t, x, i) + \sum_{j=1}^2 q_{ij} v(t, x, j) + v_x(t, x, i) r(t, i) x \\ & + \min_u \left\{ v_x(t, x, i) \rho(t, i) \sigma(t, i) u(t, x, i) + \frac{1}{2} v_{xx}(t, x, i) \sigma^2(t, i) u^2(t, x, i) \right\} = 0. \end{aligned} \quad (2.8)$$

Solving this HJB for the optimal control $u^*(\cdot)$ (i.e., investment policy) gives

$$u^*(t, x, i) = -\frac{\rho(t, i) v_x(t, x, i)}{\sigma(t, i) v_{xx}(t, x, i)}. \quad (2.9)$$

Substituting this optimal control $u^*(\cdot)$ back into (2.8), Zhou and Yin (2003) solved for the corresponding optimal value function from the HJB equation and derive the explicit form of $u^*(\cdot)$, which we summarize in the following theorem.

Theorem 2.2.1 *Problem 2.2.2 has an optimal control*

$$u^*(t, x, i) = -\frac{\rho(t, i)}{\sigma(t, i)}(x + (\lambda - z)H(t, i)) \quad (2.10)$$

and the corresponding value function is given by

$$\begin{aligned} & \inf_{u \in \mathcal{A}} J(u.; x_0, i_0, \lambda) \\ &= \mathbb{E}[(X_T^{u*} + \lambda - z)^2 | X_0^u = x_0, \alpha_0 = i_0] \\ &= V(0, x_0, i_0) \\ &= P(0, i_0)[x_0 + (\lambda - z)H(0, i_0)]^2 \\ &+ (\lambda - z)^2 \mathbb{E} \left[\int_0^T \sum_{j=1}^2 q_{\alpha(s)j} P(s, j) (H(s, j) - H(s, \alpha(s)))^2 ds \middle| \alpha_0 = i_0 \right] - \lambda^2, \end{aligned} \quad (2.11)$$

where $P(t, i)$ and $H(t, i)$ are the solutions to the following two Ordinary Differential Equations (ODEs)

$$\begin{cases} \dot{P}(t, i) = (\rho^2(t, i) - 2r(t, i))P(t, i) - \sum_{j=1}^l q_{ij}P(t, j) \\ P(T, i) = 1, \text{ for } i = 1, 2, \dots, l \end{cases} \quad (2.12)$$

$$\begin{cases} \dot{H}(t, i) = r(t, i)H(t, i) - \frac{1}{P(t, i)} \sum_{j=1}^l q_{ij}P(t, j)(H(t, j) - H(t, i)) \\ H(T, i) = 1, \text{ for } i = 1, 2, \dots, l. \end{cases} \quad (2.13)$$

While Zhou and Yin (2003) considered the market regime dynamics that affect investors' trading activity, the optimal control $u^*(t, x, i)$ (2.10) is an $\mathbb{R}$ -valued, deterministic function given the time t , portfolio value x and regime i . This indicates a precommitted policy, which is fixed and does not explore within the control space once it is considered to be optimized. However, recent work has demonstrated that a RL-facilitated exploration within the control space can achieve better performance, although under a non-Regime-Switching setting (H. Wang & Zhou, 2020). This motivates us to consider a RL-extension to Problem 2.2.2.

2.3 Exploratory Mean-Variance with Regime-Switching Problem

In this section, we make an RL extension to the aforementioned MVRS problem (Problem 2.2.2) in (Zhou & Yin, 2003), following the exploratory policy formulation from (H. Wang

& Zhou, 2020). We hereafter refer our problem to the *Exploratory Mean-Variance with Regime-Switching (EMVRS) problem*. This is a generalized problem of the past works (Zhou & Yin, 2003; H. Wang & Zhou, 2020), which accounts for the dynamics of the market regimes and allows for informed exploration within the control space.

We adopt the stock and bond dynamics, (2.1) and (2.2), and the Markov-modulated regime-switching setting. Further, in order to allow for exploration within the control space, we define the exploratory control as a probability distribution, called the *policy distribution*, in the definition below. For any subset of real numbers, $\mathcal{V} \subseteq \mathbb{R}$, we denote $\mathcal{P}(\mathcal{V})$ as the set of probability density functions defined over $\mathcal{V}$.

Definition 2.3.1 (Policy Distribution) *The policy distribution,³ $\pi := \{\pi_{t,i}(\cdot)\}_{(t,i) \in [0,T] \times \{1,\dots,l\}}$, is a distribution-valued, $\{\mathcal{F}_t\}_{t \in [0,T]}$ -adapted random variable, where each $\pi_{t,i}(\cdot) : \mathcal{U} \mapsto \mathcal{P}(\mathcal{U})$ is a probability density function over the feasible set of control values $\mathcal{U} \subseteq \mathbb{R}$ conditioned on the regime $i \in \{1, \dots, l\}$, satisfying $\int_{\mathcal{U}} \pi_{t,i}(u) du = 1$ and $\pi_{t,i}(u) \geq 0, \forall t \in [0, T], u \in \mathcal{U}$. To avoid ambiguity, we call $\pi_{t,i}$ the policy distribution at time $t \in [0, T]$.*

Remark 2.3.1 *The feasibility of control values depends on market regulations and institutional constraints. For example, if short selling is forbidden, then $u \geq 0$ which implies $\mathcal{U} \equiv \mathbb{R}_+$.*

Then, the portfolio value process is extended from (2.3) to

$$\begin{aligned} dX_t^\pi = & \left[r(t, \alpha_t) X_t^\pi + \int_{\mathcal{A}} \rho(t, \alpha_t) \sigma(t, \alpha_t) \cdot u \cdot \pi_{t, \alpha_t}(u) du \right] dt \\ & + \left(\sqrt{\int_{\mathcal{A}} \sigma^2(t, \alpha_t) u^2 \cdot \pi_{t, \alpha_t}(u) du} \right) dW_t. \end{aligned} \quad (2.14)$$

We include a brief derivation of this equation in Appendix A.1.1.

Remark 2.3.2 *We note also from the recent literature that the exploratory extension is not a trivial control randomization, as the continuum sampling of actions suffers from measurability issues; see arguments in (Jia & Zhou, 2025). Instead, we consider an augmentation of the filtration that supports independent copies of a standard Uniform random variable,*

³H. Wang and Zhou (2020) defined this in a similar way and called it the *exploratory control*.

namely $V_i \stackrel{iid}{\sim} UNIF([0, 1])$, for $i = 1, 2, \dots$. In a general set up with control $u := \{u_t\}_{t \in [0, T]}$ and state $X := \{X_t\}_{t \in [0, T]}$, the controlled state process

$$dX_t^u = b(t, X_t^u, u_t)dt + \varsigma(t, X_t^u, u_t)dW_t$$

is first discretized to a time grid $\mathbb{S} := \{0 = t_0 < t_1 < \dots < t_N = T\}$, where the control randomization takes place only at the grid points. For instance, an action randomized at t_i is held fixed over the subsequent time grid for $t \in (t_i, t_{i+1}]$. Note that in this chapter the drift coefficient $b(t, x, u) = r(t, \alpha_t)x + (\mu(t, \alpha_t) - r(t, \alpha_t))u$ and the diffusion coefficient $\varsigma(t, x, u) = \sigma(t, \alpha_t)u$, according to (2.3). The discretized controlled state process, termed as the grid sampled state process in (Gao et al., 2024), is given by

$$dX_t^{\pi, \mathbb{S}} = b(t, X_t^{\pi, \mathbb{S}}, G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, V_{i+1}))dt + \varsigma(t, X_t^{\pi, \mathbb{S}}, G^\pi(t_i, X_{t_i}^{\pi, \mathbb{S}}, V_{i+1}))dW_t,$$

where $G^d : [0, 1] \mapsto \mathbb{R}$ is the quantile function of a given distribution d . Then, as shown by (Bender & Thuan, 2026), the limit dynamics of the grid sampled state process, as the mesh size of the time grid diminishes to zero and as sampling infinitely many randomized actions at each grid point, has the same law as the exploratory state process

$$dX_t^\pi = \int_{\mathcal{A}} b(t, X_t^\pi, u)\pi_{t, X_t^\pi}(u)du dt + \sqrt{\int_{\mathcal{A}} \varsigma^2(t, X_t^\pi, u)\pi_{t, X_t^\pi}(u)du} dW_t,$$

where $\mathcal{A}$ is the action space. In this thesis, we acknowledge and adopt these exploratory extension steps, despite we do not readdress it in each of the following chapters to avoid redundancy.

Definition 2.3.2 (Admissible Policy Distribution) For any time $t \in [0, T]$ and regime $i \in \{1, \dots, l\}$, we say that $\pi_{t,i}$ is an admissible policy distribution, denoted as $\pi_{t,i} \in \mathcal{A}^\pi$, if the following conditions are satisfied:

- (i) For $t \in [0, T]$ and $i \in \{1, \dots, l\}$, $\pi_{t,i}(\cdot)$ is a policy distribution at t as described in Definition 2.3.1.
- (ii) For any feasible set of control values $\mathcal{U} \subseteq \mathbb{R}$, the stochastic process $\{\int_{\mathcal{U}} \pi_{t,i}(u)du\}_{(t,i) \in [0, T] \times \{1, \dots, l\}}$ is $\{\mathcal{F}_t\}_{t \in [0, T]}$ -adapted.
- (iii) For all $(t, i) \in [0, T] \times \{1, \dots, l\}$,

$$\mathbb{E} \left[\int_t^T \int_{\mathcal{A}} u^2 \pi_{s, \alpha_s}(u) du ds \middle| \alpha_t = i \right] < \infty. \quad (2.15)$$

This implies that the exploratory wealth dynamic SDE (2.14) has a unique solution X_t^π corresponding to $\pi_{t,i}(\cdot)$, according to the arguments in (Zhou & Yin, 2003; Dai et al., 2023).

(iv) For all $(t, x, i) \in [0, T] \times \mathbb{R} \times \{1, \dots, l\}$ and constant $0 < \xi < \infty$,

$$\mathbb{E} \left[(X_T^\pi + \lambda - z)^2 + \xi \int_t^T \int_{\mathcal{A}} \pi_{s, \alpha_s}(u) \log \pi_{s, \alpha_s}(u) du ds \middle| X_t^\pi = x, \alpha_t = i \right] < \infty. \quad (2.16)$$

If $\pi_{t,i} \in \mathcal{A}^\pi, \forall (t, i) \in [0, T] \times \{1, \dots, l\}$, we denote $\pi \in \mathcal{A}^\pi$.

We propose the EMVRS problem as an entropy-regularized version of the MVRS problem (2.2.2), which restrains the policy distribution within a certain family of distributions.

Problem 2.3.1 (The EMVRS Problem)

$$\min_{\pi \in \mathcal{A}^\pi} \mathbb{E} \left[(X_T^\pi + \lambda - z)^2 + \xi \int_0^T \int_{\mathcal{A}} \pi_{t, \alpha_t}(u) \log \pi_{t, \alpha_t}(u) du dt \middle| X_0^\pi = x_0, \alpha_0 = i_0 \right] - \lambda^2, \quad (2.17)$$

where $\xi > 0$ is the exploration weight, $x_0 > 0$ and $i_0 \in \{1, \dots, l\}$ are respectively the portfolio value and market regime at $t = 0$.

For any $(t, x) \in [0, T] \times \mathbb{R}, i \in \{1, \dots, l\}$ and admissible policy distribution $\pi \in \mathcal{A}^\pi$, we define the *value function* (corresponding to π) V^π and the *optimal value function* V^* as

$$V^\pi(t, x, i) := \mathbb{E} \left[(X_T^\pi + \lambda - z)^2 + \xi \int_t^T \int_{\mathcal{A}} \pi_{k, \alpha_k}(u) \log \pi_{k, \alpha_k}(u) du dk \middle| X_t^\pi = x, \alpha_t = i \right] - \lambda^2 \quad (2.18)$$

$$V^*(t, x, i) := \inf_{\pi \in \mathcal{A}^\pi} V^\pi(t, x, i). \quad (2.19)$$

Following standard DPP arguments and assuming the value function is smooth, we derive the HJB equation

$$\begin{aligned} & v_t(t, x, i) + v_x(t, x, i)r(t, i)x + \sum_{j=1}^l q_{ij}v(t, x, j) \\ & - \frac{1}{2} \frac{\rho^2(t, i)v_x^2(t, x, i)}{v_{xx}(t, x, i)} - \frac{\xi}{2} \log \left(\frac{2\pi\xi}{\sigma^2(t, i)v_{xx}(t, x, i)} \right) = 0. \end{aligned} \quad (2.20)$$

The derivation of (2.20) is given in Appendix A.1.2.

Then, we solve this HJB equation (2.20) for the optimal value function V^* , which is given in the theorem below. The proof is postponed to Appendix A.1.3.

Theorem 2.3.1 (Solution to the EMVRS Problem) *Problem 2.3.1 has an optimal policy distribution $\pi^* = \{\pi_t^*\}_{t \in [0, T]}$, where each π_t^* is given by a Gaussian distribution*

$$\pi_{t,i}^*(u) = N\left(-\frac{\rho(t,i)}{\sigma(t,i)}[x + (\lambda - z)H(t,i)], \frac{\xi}{2\sigma^2(t,i)P(t,i)}\right), \quad (2.21)$$

with $(t, x) \in [0, T] \times \mathbb{R}$ and $i \in \{1, \dots, l\}$. The corresponding optimal value function is given by

$$V^*(t, x, i) = P(t, i)[x + (\lambda - z)H(t, i)]^2 + (\lambda - z)^2 C(t, i) + D(t, i) - \lambda^2, \quad (2.22)$$

where $P(t, i), H(t, i)$ are solutions of the two ODEs in (2.12) and (2.13), and

$$C(t, i) = \sum_{m=1}^l \sum_{j=1}^l \int_t^T p_{im}(s-t) q_{mj} P(s, j) (H(s, j) - H(s, m))^2 ds, \quad (2.23)$$

$$D(t, i) = - \sum_{m=1}^l \int_t^T p_{im}(s-t) \frac{\xi}{2} \log\left(\frac{\pi e \xi}{\sigma^2(s, m) P(s, m)}\right) ds. \quad (2.24)$$

Moreover, at initialization with $t = 0$, the optimal Lagrange multiplier is

$$\lambda^* = \frac{z - P(0, i_0)H(0, i_0)x_0}{P(0, i_0)H(0, i_0)^2 + C(0, i_0) - 1} + z. \quad (2.25)$$

Remark 2.3.3 *The functions $C(t, i)$ and $D(t, i)$ are solutions to the following two ODEs*

$$\begin{cases} \dot{C}(t, i) = - \sum_{j=1}^l q_{ij} [P(t, j)(H(t, j) - H(t, i))^2 + C(t, j)], \\ C(T, i) = 0, \text{ for } i \in \{1, \dots, l\}, \end{cases} \quad (2.26)$$

$$\begin{cases} \dot{D}(t, i) = \frac{\xi}{2} \log\left(\frac{\pi \xi}{\sigma^2(t, i) P(t, i)}\right) - \sum_{j=1}^l q_{ij} D(t, j), \\ D(T, i) = 0, \text{ for } i \in \{1, \dots, l\}. \end{cases} \quad (2.27)$$

Remark 2.3.4 *When there is only one regime, the market parameters $\mu(t, \alpha_t), \sigma(t, \alpha_t)$ and $r(t, \alpha_t)$ in (2.1) and (2.2) collapse to $\mu(t), \sigma(t), r(t)$ respectively. H. Wang and Zhou*

(2020) further set them to constants ($\mu(t) = \mu, \sigma(t) = \sigma, r(t) = r$) that do not vary with time, which simplifies Problem 2.3.1 to the EMV problem

$$\min_{\pi \in \mathbb{P}(\mathcal{A})} \mathbb{E} \left[(X_T^\pi + \lambda - z)^2 + \xi \int_0^T \int_{\mathcal{A}} \pi_t(u) \log \pi_t(u) du dt \middle| X_0^\pi = x_0 \right] - \lambda^2, \quad (2.28)$$

where $\xi > 0$ is the exploration weight. The EMVRS solution in Theorem 2.3.1 reduces to the following corollary. The proof is given in Appendix A.1.4.

Corollary 2.3.1 (Solution to the EMV Problem) *The EMV problem has an optimal policy distribution $\pi^* = \{\pi_t^*(\cdot)\}_{t \in [0, T]}$, where each $\pi_t^* : \mathcal{A} \mapsto \mathbb{P}(\mathcal{A})$ is given by the Gaussian distribution*

$$\pi_t^*(u) = N \left(u \middle| -\frac{\rho}{\sigma} (x + (\lambda - z)e^{-r(T-t)}), \frac{\xi}{2\sigma^2} e^{(\rho^2 - 2r)(T-t)} \right) \quad (2.29)$$

with $\lambda^* = z - \frac{ze^{(\rho^2 - r)(T-t)} - x_0}{e^{\rho^2 T} - 1}$. The corresponding value function is given by

$$\begin{aligned} v(t, x, \lambda) &= (x + (\lambda - z)e^{-r(T-t)})^2 e^{-(\rho^2 - 2r)(T-t)} \\ &\quad + \frac{\xi(\rho^2 - 2r)}{4} (T^2 - t^2) - \frac{\xi}{2} \left[(\rho^2 - 2r)T - \log \frac{\sigma^2}{\pi\xi} \right] (T - t) - \lambda^2. \end{aligned} \quad (2.30)$$

Finally, Theorem 2.3.1 provides the optimal policy distribution that solves the EMVRS problem. However, in practice we usually start with an initial guess of the policy π^0 and iteratively update it, until it converges (or is sufficiently close) to the optimal solution π^* . In the RL literature, this is known as *policy iteration* (Sutton & Barto, 2018). In the following Policy Improvement Theorem (PIT), we propose an iteration scheme for policy updates, which is guaranteed to improve or at least not downgrade the current policy. The proof is given in Appendix A.1.5.

Theorem 2.3.2 (Policy Improvement Theorem) *Let $\pi \in \mathcal{A}^\pi$ be any admissible policy distribution and $V^\pi(\cdot, \cdot, i)$ be the corresponding value function as defined in (2.18), satisfying $V_{xx}^\pi(t, x, i) > 0$ for any regime $i \in \{1, \dots, l\}$, time and wealth $(t, x) \in [0, T] \times \mathbb{R}$. Consider constructing a new policy distribution $\pi^* = \{\pi_{t,i}^*\}_{(t,i) \in [0, T] \times \{1, \dots, l\}}$, where each $\pi_{t,i}^*$ is given by*

$$\pi_{t,i}^*(u) = N \left(-\frac{\rho(t, i)V_x^\pi(t, x, i)}{\sigma(t, i)V_{xx}^\pi(t, x, i)}, \frac{\xi}{\sigma^2(t, i)V_{xx}^\pi(t, x, i)} \right). \quad (2.31)$$

If this new policy is admissible, i.e., $\pi^* \in \mathcal{A}^\pi$, then $V^{\pi^*}(t, x, i) \leq V^\pi(t, x, i)$, for all $(t, x) \in [0, T] \times \mathbb{R}, i \in \{1, \dots, l\}$.

According to the PIT, we can always upgrade, or at least not downgrade, the investment policy given the value function. So, the optimality of the policy relies on the optimality of the corresponding value function, which we discuss in more detail in Section 2.4.

2.4 RL Algorithm

To optimize the value function, we develop an RL algorithm with the policy updating scheme following the PIT (Theorem 2.3.2). We realize that the PIT constructs an improved policy distribution based on the *market parameters*, stock return volatility $\sigma(t, i)$ and Sharpe Ratio $\rho(t, i)$ at time t and regime i , which are inaccessible in practice. Meanwhile, the optimal value function defined in Theorem 2.3.1 also depends on the market parameters, although indirectly through the P, H, C, D functions. Thus, the optimization of the value function is a matter of learning the market parameters. In other words, if we did know the “true” market parameters, we could turn off policy exploration (i.e., setting $\xi = 0$) and adopt the optimal investment strategy of the classical MV problem in (2.9).

For ease of presentation, we hereafter suppose that there are only two regimes, i.e., $l = 2$ and $\{\alpha_t\}_{t \in [0, T]} \in \{1, 2\}$. We further assume that the market parameters and the interest rates are constant in the same regime, regardless of the time. That is, for any $t_1, t_2 \in [0, T], i \in \{1, 2\}$

$$\sigma(t_1, i) = \sigma(t_2, i) =: \sigma_i; \quad \rho(t_1, i) = \rho(t_2, i) =: \rho_i; \quad r(t_1, i) = r(t_2, i) = r_i.$$

So, the value function can be seen as a function of $\theta := (\sigma_1, \sigma_2, \rho_1, \rho_2)$,

$$V^\theta(t, x, i) := P^\theta(t, i)[x + (\lambda - z)H^\theta(t, i)]^2 + (\lambda - z)^2 C^\theta(t, i) + D^\theta(t, i) - \lambda^2, \quad (2.32)$$

where $P^\theta, H^\theta, C^\theta, D^\theta$ are the solution to the system of ODEs in (2.12), (2.13), (2.26) and (2.27). We use the superscript θ to denote the direct dependence of P, H, C, D on θ and the indirect dependence of V^* on θ . Denote $\{X_t^\theta\}_{t \in [0, T]}$ as the portfolio value process that follows the investment policy π^θ . Then, the PIT-inferred policy distribution $\pi^\theta := \{\pi_{t,i}^\theta\}_{(t,i) \in [0, T] \times \{1, \dots, l\}}$ is reparameterized as

$$\begin{aligned} \pi_{t,i}^\theta(u) &:= N\left(u \left| -\frac{\rho_i V_x^\theta(t, x, i)}{\sigma_i V_{xx}^\theta(t, x, i)}, \frac{\xi}{\sigma_i^2 V_{xx}^\theta(t, x, i)} \right. \right) \\ &= N\left(u \left| -\frac{\rho_i}{\sigma_i} [x + (\lambda - z)H^\theta(t, i)]^2, \frac{\xi}{2\sigma_i^2 P^\theta(t, i)} \right. \right), \end{aligned} \quad (2.33)$$

given $X_t^\theta = x \in \mathbb{R}$ and $\alpha_t = i \in \{1, 2\}$.

In the rest of this section, we will introduce two updating schemes for the market parameters — the Temporal Difference (TD) Learning (Section 2.4.1) and the Orthogonality Condition Learning (Section 2.4.2). The TD learning was originally introduced by (Sutton, 1988), and later H. Wang and Zhou (2020) applied it for the RL training of their EMV problem. More recently, Jia and Zhou (2022a) argued against the appropriateness of TD learning in stochastic control problems and proposed to leverage the martingale property of the value function in the RL training process. We wrap them up with a training algorithm and give the corresponding updating scheme for the market parameters under a simulated market setting in Section 2.4.3. We show both theoretically and empirically (in the next section) that the TD Learning is not suitable for the EMVRS problem. Finally, we conclude this section with some remarks on how our algorithm can be amended to train EMVRS on real market data in Section 2.4.4.

2.4.1 Temporal Difference (TD) Learning

We know by Bellman's Principle of Optimality that the value function has the recursive form, for $0 \leq t \leq s \leq T, x \in \mathbb{R}, i \in \{1, 2\}$

$$V^*(t, x, i) = \mathbb{E} \left[V^*(s, X_s^\pi, \alpha_s) + \xi \int_t^s \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk \middle| X_t^\pi = x, \alpha_t = i \right] \quad (2.34)$$

$$= \mathbb{E} \left[V^*(s, X_s^\pi, \alpha_s) - \xi \int_t^s \frac{1}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_k}^2 P(t, \alpha_k)} \right) dk \middle| X_t^\pi = x, \alpha_t = i \right], \quad (2.35)$$

where the second equality is a direct substitution of the entropy of the Gaussian distribution π_{k, α_k}^* . This further gives an expectation of the *temporal difference* in the value function from t to s :

$$\mathbb{E} \left[\frac{V^*(s, X_s^\pi, \alpha_s) - V^*(t, x, i)}{s - t} - \frac{\xi}{s - t} \int_t^s \frac{1}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_k}^2 P(t, \alpha_k)} \right) dk \middle| X_t^\pi = x, \alpha_t = i \right] = 0.$$

Taking $s \downarrow t$, we can intuitively treat the left-hand-side as the expectation of an instantaneous improvement of the value function from time t , following the investment policy $\pi_{t, i}^*$. Such an expectation being zero at all $t \in [0, T]$ is considered ideal because this implies no further improvement is attainable. While this expectation is intractable in the continuous time setting, Temporal Difference (TD) Learning allows us to approximate this expectation by its discretized counterpart.

Consider a discretization of the investment horizon $0 = t_0 < \dots t_k < \dots t_K = T$, with mesh size equal to $\Delta t := t_{k+1} - t_k$ and $K := \frac{T}{\Delta t}$. We define the *TD loss function* as the mean-squared sum of the temporal differences measured at $\{t_k\}_{k=0, \dots, K}$

$$TD(\theta) := \frac{1}{2} \mathbb{E} \left[\sum_{k=0}^{K-1} \left(\frac{V^\theta(t_{k+1}, X_{t_{k+1}}^\theta, \alpha_{t_{k+1}}) - V^\theta(t_k, X_{t_k}^\theta, \alpha_{t_k})}{\Delta t} - \frac{\xi}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_{t_k}}^2 P^\theta(t_k, \alpha_{t_k})} \right) \right)^2 \Delta t \right]. \quad (2.36)$$

The above expectation is taken over the filtered probability space of the market and the policy exploration, which can be approximated by simulating trajectories of portfolio values $\{X_{t_k}^\theta\}_{k=1}^K$ and the Markovian regimes $\{\alpha_{t_k}\}_{k=1}^K$. We use Stochastic Gradient Descent (SGD) to minimize the TD loss, which we describe in more detail in Section 2.4.3.

Here, we remind the readers that the recursive form (2.34) only holds for the optimal value function V^* , which is supposed to be attained when the market parameters θ converge to the “grounding true” values θ_{true} . On the other hand, if we define a process

$$M_t^* := V^*(t, x, i) + \int_0^t \xi \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk \quad (2.37)$$

given $X_t^\pi = x \in \mathbb{R}$ and $\alpha_t = i \in \{1, 2\}$, then $M^* := \{M_t^*\}_{t \in [0, T]}$ is a martingale according to (2.34), with the following dynamics

$$\begin{aligned} dM_t^* &= dV^*(t, x, i) + \left(\xi \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du \right) dt \\ &= \left\{ V_t^*(t, x, i) + V_x^*(t, x, i) r_i x + \sum_{j=1}^2 q_{ij} V^*(t, x, i) \right. \\ &\quad \left. - \frac{1}{2} \frac{\rho_i^2 (V_x^*(t, x, i))^2}{V_{xx}^*(t, x, i)} - \frac{\xi}{2} \log \left(\frac{2\pi\xi}{\sigma_i^2 V_{xx}^*(t, x, i)} \right) \right\} dt \\ &\quad + \left\{ V_x^*(t, x, i) \sigma_i \sqrt{\frac{\xi}{\sigma_i^2 V_{xx}^*(t, x, i)} + \left(\frac{\rho_i V_x^*(t, x, i)}{\sigma_i V_{xx}^*(t, x, i)} \right)^2} \right\} dW_t. \end{aligned}$$

Based on the TD loss (2.36), we have:

$$\frac{1}{\Delta t} \sum_{k=0}^{K-1} \left(V^*(t_{k+1}, X_{t_{k+1}}^\pi, \alpha_{t_{k+1}}) - V^*(t_k, X_{t_k}^\pi, \alpha_{t_k}) \right)$$

$$\begin{aligned}
& -\frac{\xi}{2} \int_{t_k}^{t_{k+1}} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_s}^2 P(s, \alpha_s)} \right) ds + \mathcal{O}(\Delta t^2) \Big)^2 \\
& \approx \frac{1}{\Delta t} \langle M^* \rangle_T \\
& = \frac{1}{\Delta t} \int_0^T (V_x^*(t, X_t^\pi, \alpha_t))^2 \sigma_{\alpha_t}^2 \left(\frac{\xi}{\sigma_{\alpha_t}^2 V_{xx}^*(t, X_t^\pi, \alpha_t)} + \left(\frac{\rho_{\alpha_t}^2 (V_x^*(t, X_t^\pi, \alpha_t))^2}{\sigma_{\alpha_t}^2 (V_{xx}^*(t, X_t^\pi, \alpha_t))^2} \right)^2 \right) dt.
\end{aligned}$$

This means that training with the TD loss is equivalent to minimizing the quadratic variation of M^* , which is not zero and should not be minimized either. This implies that minimizing the TD loss is inadequate for the EMVRS problem. This agrees with the arguments in (Jia & Zhou, 2022a), although H. Wang and Zhou (2020) used a similar TD loss for the EMV problem. This motivates us to proceed with another method to learn the market parameters.

2.4.2 Orthogonality Condition (OC) Learning

For some market parameters θ and their “grounding truth” $\theta_{true} := (\sigma_{true,1}, \sigma_{true,2}, \rho_{true,1}, \rho_{true,2})$, we define

$$\begin{aligned}
M_t^\theta &:= V^\theta(t, X_t^\theta, \alpha_t) + \int_0^t \xi \int_{\mathcal{A}} \pi_{k, \alpha_k}^\theta(u) \log \pi_{k, \alpha_k}^\theta(u) du dk \\
&= V^\theta(t, X_t^\theta, \alpha_t) - \frac{\xi}{2} \int_0^t \log \left(\frac{\pi e \xi}{\sigma_{\alpha_k}^2 P^\theta(k, \alpha_k)} \right) dk.
\end{aligned} \tag{2.38}$$

We remind the readers that the portfolio value is derived from the market dynamics, which should use θ_{true} . So, the portfolio value process is rewritten from (2.14) to

$$\begin{aligned}
dX_t^\theta &= [r_{\alpha_t} X_t^\theta + \int_{\mathcal{A}} \rho_{true, \alpha_t} \sigma_{true, \alpha_t} \cdot u \cdot \pi_{t, \alpha_t}^\theta(u) du] dt + \sqrt{\int_{\mathcal{A}} \sigma_{true, \alpha_t}^2 u^2 \pi_{t, \alpha_t}^\theta(u) du} dW_t \tag{2.39} \\
&= \left\{ r_{\alpha_t} X_t^\theta + \rho_{true, \alpha_t} \sigma_{true, \alpha_t} \left(-\frac{\rho_{\alpha_t}}{\sigma_{\alpha_t}} [X_t^\theta + (\lambda - z) H^\theta(t, \alpha_t)] \right) \right\} dt \\
&+ \left\{ \sigma_{true, \alpha_t} \sqrt{\frac{\rho_{\alpha_t}^2}{\sigma_{\alpha_t}^2} [X_t^\theta + (\lambda - z) H^\theta(t, \alpha_t)]^2 + \frac{\xi}{2 \sigma_{\alpha_t}^2 P^\theta(t, \alpha_t)}} \right\} dW_t.
\end{aligned} \tag{2.40}$$

Applying Itô's Formula and after some calculations, we get the drift term of dM^θ

$$\begin{aligned}
& \left\{ \dot{P}^\theta(t, \alpha_t) + P^\theta(t, \alpha_t) \sigma_{true, \alpha_t}^2 \frac{\rho_{\alpha_t}^2}{\sigma_{\alpha_t}^2} - 2P^\theta(t, \alpha_t) \rho_{true, \alpha_t} \sigma_{true, \alpha_t} \frac{\rho_{\alpha_t}}{\sigma_{\alpha_t}} \right. \\
& \quad \left. + 2P^\theta(t, \alpha_t) r_{\alpha_t} + \sum_{j=1}^l q_{\alpha_t j} P^\theta(t, j) \right\} \times [X_t + (\lambda - z) H^\theta(t, \alpha_t)]^2 \\
& + \left\{ \dot{H}^\theta(t, \alpha_t) - r_{\alpha_t} H^\theta(t, \alpha_t) + \frac{1}{P^\theta(t, \alpha_t)} \sum_{j=1}^l q_{\alpha_t j} P^\theta(t, j) [H^\theta(t, j) - H^\theta(t, \alpha_t)] \right\} \\
& \quad \times 2(\lambda - z) P^\theta(t, \alpha_t) [X_t + (\lambda - z) H^\theta(t, \alpha_t)] \\
& + \left\{ \dot{C}^\theta(t, \alpha_t) + \sum_{j=1}^l q_{\alpha_t j} [P^\theta(t, j) [H^\theta(t, j) - H^\theta(t, \alpha_t)]^2 + C^\theta(t, j)] \right\} \times (\lambda - z)^2 \\
& + \left\{ \dot{D}^\theta(t, \alpha_t) + \sum_{j=1}^l q_{\alpha_t j} D^\theta(t, j) - \frac{\lambda}{2} \log \left(\frac{\pi \lambda}{\sigma_{\alpha_t}^2 P^\theta(t, \alpha_t)} \right) \right\}.
\end{aligned}$$

This equals zero if and only if $\theta = \theta_{true}$, given that $P^\theta, H^\theta, C^\theta, D^\theta$ are the solutions to the ODEs in (2.12), (2.13), (2.26) and (2.27). That is, M_t^θ is a martingale if and only if θ coincides with θ_{true} .

Moreover, we know that any square-integrable martingale $M := \{M_t\}_{t \in [0, T]}$ has an orthogonality condition

$$\mathbb{E} \left[\int_0^T \zeta_t dM_t \right] = 0, \quad (2.41)$$

where the *test function* $\zeta := \{\zeta_t\}_{t \in [0, T]}$ is any $\{\mathcal{F}_t\}$ -adapted process that is square-integrable with respect to M . According to (Jia & Zhou, 2022a), this is a necessary and sufficient condition to characterize the martingality of M , regardless of the choice of the test function. Hence, we define the *Orthogonality Condition (OC) loss* by taking ζ as the partial derivative of the value function with respect to the market parameters and discretizing the continuous time.

Definition 2.4.1 (Orthogonality Condition (OC) loss) Let $\theta \equiv (\theta_1, \theta_2, \theta_3, \theta_4) \equiv (\sigma_1, \sigma_2, \rho_1, \rho_2)$

and for $j = 1, \dots, 4$,

$$\begin{aligned}
OC(\theta_j) &= \mathbb{E} \left[\sum_{k=0}^{K-1} \frac{\partial V^\theta(t_k, X_{t_k}, \alpha_{t_k})}{\partial \theta_j} (M_{t_{k+1}}^\theta - M_{t_k}^\theta) \right] \\
&= \mathbb{E} \left[\sum_{k=0}^{K-1} \frac{\partial V^\theta(t_k, X_{t_k}, \alpha_{t_k})}{\partial \theta_j} \left(V^\theta(t_{k+1}, X_{t_{k+1}}^\theta, \alpha_{t_{k+1}}) - V^\theta(t_k, X_{t_k}^\theta, \alpha_{t_k}) \right. \right. \\
&\quad \left. \left. - \frac{\xi}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_{t_k}}^2 P^\theta(t_k, \alpha_{t_k})} \right) \Delta t \right) \right]. \tag{2.42}
\end{aligned}$$

Again, the above expectation is taken over the filtered probability space of the market and the policy exploration, which can be approximated by simulating trajectories of portfolio values $\{X_{t_k}^\theta\}_{k=1}^K$ and the Markovian regimes $\{\alpha_{t_k}\}_{k=1}^K$.

2.4.3 Updating Scheme and Training Algorithm

We note that the randomness in our problem comes from three sources — the market dynamics, the Markovian regimes and the policy exploration. Therefore, it is impractical to exactly compute the TD loss (2.36) or the OC loss (2.42), as they are defined to be expectations. Instead, we adopt batch-training over a predetermined number of epochs N_{epochs} , which is set to be large enough for market parameters to converge.

For each epoch $n = 1, \dots, N_{epochs}$, we fix a realized path of Brownian motion $\{W_{t_k}^{(n)}\}_{k=0}^K$, where each $\Delta W_{t_k}^{(n)} := W_{t_{k+1}}^{(n)} - W_{t_k}^{(n)}$ follows a Gaussian distribution with zero mean and variance Δt . We hereafter use the superscript (n) to denote a realized path of a process that we fix in epoch n . This helps us eliminate the randomness of the market dynamics. Then, we simulate a path of market regimes $\{\alpha_{t_k}^{(n)}\}_{k=0}^K$ following the categorical distribution below, given a randomly selected initial regime $\alpha_{t_0}^{(n)} \in \{1, 2\}$ and a predefined Markov Chain generator Q .

$$\alpha_{t_{k+1}}^{(n)} = \begin{cases} 1, & \text{with probability } P(\alpha_{t_{k+1}}^{(n)} = 1 | \alpha_{t_k}^{(n)}) = p_{\alpha_{t_k}^{(n)} 1} \\ 2, & \text{with probability } P(\alpha_{t_{k+1}}^{(n)} = 2 | \alpha_{t_k}^{(n)}) = p_{\alpha_{t_k}^{(n)} 2} \end{cases} \tag{2.43}$$

Here $(p_{ij})_{i=1,2}^{j=1,2} =: P$ is the transition matrix of the Markov Chain, which is the matrix exponential of the Markov Chain generator Q , i.e., $P = e^Q$. This eliminates the randomness of the Markovian regimes in epoch n . Finally, we use the “grounding true” market

parameters θ_{true} and the current market parameters $\theta^{(n)}$ to compute a path of the portfolio values $\{X_{t_k}^{(n)}\}_{k=0}^K$, given a predetermined initial portfolio value $X_{t_0}^{(n)} = x_0 > 0$

$$X_{t_{k+1}}^{(n)} = X_{t_k}^{(n)} + \Delta X_{t_k}^{(n)}, \text{ for } k = 1, \dots, K-1 \quad (2.44)$$

where $\Delta X_{t_k}^{(n)}$ is a discretized version of (2.40)

$$\begin{aligned} \Delta X_{t_k}^{(n)} &= \left\{ r_{\alpha_{t_k}^{(n)}} X_{t_k}^{(n)} + \rho_{true, \alpha_{t_k}^{(n)}} \sigma_{true, \alpha_{t_k}^{(n)}} \left(-\frac{\rho_{\alpha_{t_k}^{(n)}}}{\sigma_{\alpha_{t_k}^{(n)}}} \left[X_{t_k}^{(n)} + (\lambda - z) H^\theta(t_k, \alpha_{t_k}^{(n)}) \right] \right) \right\} \Delta t \\ &+ \left\{ \sigma_{true, \alpha_{t_k}^{(n)}} \sqrt{\frac{\rho_{\alpha_{t_k}^{(n)}}^2}{\sigma_{\alpha_{t_k}^{(n)}}^2} \left[X_{t_k}^{(n)} + (\lambda - z) H^\theta(t_k, \alpha_{t_k}^{(n)}) \right]^2 + \frac{\xi}{2\sigma_{\alpha_{t_k}^{(n)}}^2} P^\theta(t_k, \alpha_{t_k}^{(n)})} \right\} \Delta W_{t_k}^{(n)}. \end{aligned}$$

This helps us eliminate the stochasticity of policy exploration within an epoch. We hereby emphasize the hybrid usage of θ_{true} and $\theta^{(n)}$ when simulating paths of portfolio value. The θ_{true} , i.e., $(\rho_{true}, \sigma_{true})$, in the equation above was directly derived from the stock dynamics in (2.1). This is supposed to be the “grounding true” parameters because the stock price is driven by the “true” market model, despite that we cannot observe the parameters of it. This is the only place where we used θ_{true} in the algorithm.

Up to this point, we have collected two paths, the portfolio values $\{X_{t_k}^{(n)}\}_{k=0}^K$ and market regimes $\{\alpha_{t_k}^{(n)}\}_{k=0}^K$, with which we can solve for $\{P^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$, $\{H^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$, $\{C^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$ and $\{D^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$ from the system of four ODEs in (2.12), (2.13), (2.26) and (2.27). We can thereby compute the value function $\{V_{t_k}^{\theta, (n)}\}_{k=0}^K$ via (2.32) and the optimal Lagrange multiplier $\lambda^{(n)}$ via a modified version of (2.25)

$$\lambda^{(n)} = \frac{z - P^{(n)}(t_0, \alpha_{t_0}^{(n)}) H^{(n)}(t_0, \alpha_{t_0}^{(n)}) x_0}{P^{(n)}(t_0, \alpha_{t_0}^{(n)}) H^{(n)}(t_0, \alpha_{t_0}^{(n)})^2 + C^{(n)}(t_0, \alpha_{t_0}^{(n)}) - 1} + z. \quad (2.45)$$

We note that it is impractical to compute the TD loss and the OC loss in (2.36) and (2.42), due to the multiple sources of randomness. So, we instead compute the *realized TD loss* and *realized OC loss*, respectively denoted by $TD(\theta^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\})$ and

$OC(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\})$, for $j = 1, \dots, 4$.

$$\begin{aligned}
& TD(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\}) \\
&= \frac{1}{2} \sum_{k=0}^{K-1} \left[\frac{V^{\theta, (n)}(t_{k+1}, X_{t_{k+1}}^{(n)}, \alpha_{t_{k+1}}^{(n)}) - V^{\theta, (n)}(t_k, X_{t_k}^{(n)}, \alpha_{t_k}^{(n)})}{\Delta t} \right. \\
&\quad \left. - \frac{\xi}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_{t_k}^{(n)}}^2 P^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})} \right) \right]^2 \Delta t
\end{aligned} \tag{2.46}$$

$$\begin{aligned}
& OC(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\}) \\
&= \sum_{k=0}^{K-1} \frac{\partial V^{\theta, (n)}(t_k, X_{t_k}^{(n)}, \alpha_{t_k}^{(n)})}{\partial \theta_j} \times \left[V^{\theta, (n)}(t_{k+1}, X_{t_{k+1}}^{(n)}, \alpha_{t_{k+1}}^{(n)}) - V^{\theta, (n)}(t_k, X_{t_k}^{(n)}, \alpha_{t_k}^{(n)}) \right. \\
&\quad \left. - \frac{\xi}{2} \log \left(\frac{\pi e \xi}{\sigma_{\alpha_{t_k}^{(n)}}^2 P^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})} \right) \Delta t \right].
\end{aligned} \tag{2.47}$$

To update the market parameters, we apply the SGD on the TD loss and the Stochastic Optimization method on the OC loss. For $j = 1, \dots, 4$, the updating scheme of θ_j with the TD loss is

$$\theta_j^{(n+1)} \leftarrow \theta_j^{(n)} - \eta_{TD,j} \frac{\partial TD(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\})}{\partial \theta_j}, \tag{2.48}$$

and the updating scheme of θ_j with the OC loss is

$$\theta_j^{(n+1)} \leftarrow \theta_j^{(n)} + \eta_{OC,j} OC(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\}), \tag{2.49}$$

where $\eta_{TD,j}, \eta_{OC,j} > 0$ are the learning rates for θ_j when using the TD loss and the OC loss respectively. In practice, the partial derivative is nontrivial and does not have an explicit form in the EMVRS problem. So, we approximate it via the central difference method

$$\frac{\partial TD(\theta_j^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\})}{\partial \theta_j} \approx \frac{TD(\theta_{j+}^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\}) - TD(\theta_{j-}^{(n)}; \{X_{t_k}^{(n)}\}, \{\alpha_{t_k}^{(n)}\})}{2\epsilon_j},$$

where $\epsilon_j > 0$ is a small step size and $\theta_{j\pm}^{(n)}$ equals $\theta^{(n)}$ except the j -th entry is replaced by $\theta_j \pm \epsilon_j$.

Finally, we summarize our method in Algorithm 1.

Algorithm 1: The Training Algorithm for EMVRS with Simulated Data

Initialize the hyperparameters: number of epochs N_{epochs} , the investment horizon T , the mesh size of the continuous time discretization Δt , the Markov Chain generator Q , the exploration parameter ξ , the initial portfolio value x_0 , the target terminal portfolio value z , the learning rates $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$, the “grounding true” market parameters $\theta_{true} = (\theta_{true,1}, \theta_{true,2}, \theta_{true,3}, \theta_{true,4}) \equiv (\sigma_{true,1}, \sigma_{true,2}, \rho_{true,3}, \rho_{true,4})$.

Initialize the market parameters (to be learned through RL training)
 $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}) \equiv (\sigma_1^{(0)}, \sigma_2^{(0)}, \rho_1^{(0)}, \rho_2^{(0)})$ and the interest rates (r_1, r_2) .

for $n = 0, \dots, N_{epochs} - 1$ **do**

- Fix a realized path of Brownian motion $\{W_{t_k}^{(n)}\}_{k=0}^K$, with $\Delta W_{t_k}^{(n)} := W_{t_{k+1}}^{(n)} - W_{t_k}^{(n)} \sim N(0, \Delta t)$.
- Simulate a path of portfolio values $\{X_{t_k}^{(n)}\}_{k=0}^K$ and regimes $\{\alpha_{t_k}^{(n)}\}_{k=0}^K$ via (2.44) and (2.43).
- Solve the ODEs in (2.12), (2.13), (2.26) and (2.27) for $\{P^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$, $\{H^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$, $\{C^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$ and $\{D^{\theta, (n)}(t_k, \alpha_{t_k}^{(n)})\}$.
- Compute $\{V_{t_k}^{\theta, (n)}\}_{k=0}^K$ via (2.32) and $\lambda^{(n)}$ via (2.45).
- Update $\theta_j^{(n)}$ via the updating scheme (2.48) if using TD learning or via (2.49) if using OC learning, for $j = 1, \dots, 4$.

end

2.4.4 Training on Real Market Data

If we can observe real market data, the source of randomness of our problem reduces to policy exploration, since the observed stock prices $\{S_{t_k}\}_{k=0}^K$ and the interest rates $\{r_{t_k}\}_{k=0}^K$ over a predetermined observation window are fixed. The market regimes $\{\alpha_{t_k}\}_{k=0}^K$, although not directly observable, can be estimated via fitting a Hidden Markov Model and implementing the Viterbi Algorithm (Viterbi, 1967). The Viterbi Algorithm outputs the maximum a posteriori estimate of the hidden state sequence, which can be taken as a surrogate for the market regimes.

The training algorithm with real data is the same as Algorithm 1, except we no longer need to sample paths of Brownian motion, nor do we have knowledge of the “ground true” market parameters. Instead, we generate a path of portfolio values $\{X_{t_k}^{(n)}\}_{k=0}^K$ by

first sampling investment action $u_{t_k}^{(n)}$ from the current policy $\pi_{t_k, \alpha_{t_k}}^{\theta^{(n)}}(\cdot)$ and then iteratively computing the next portfolio value via

$$X_{t_{k+1}}^{(n)} = u_{t_k}^{(n)} \frac{S_{t_{k+1}}}{S_{t_k}} + (X_{t_k}^{(n)} - u_{t_k}^{(n)})(1 + r_{t_k} \Delta t)$$

for $k = 0, \dots, K - 1$ and given $X_0^{(n)} = x_0 > 0$. The TD loss and OC loss are defined in the same way as in (2.36) and (2.42), except the expectations are taken over the filtered probability space of policy exploration only and are conditioned on the observed real market data.

2.5 Numerical Results

In this section, we demonstrate the performance of our proposed method with both simulated data and real market data. Two key lessons emerge from the results. First, OC learning leads to the parameters' convergence to their corresponding “grounding truth” in the simulation study, where the “grounding truths” are given upon initialization. However, TD learning does not guarantee such convergence. Second, EMVRS outperforms EMV on the real market data under different investment constraint settings, achieving higher mean terminal portfolio value and relatively lower volatility.

2.5.1 Simulation Study: Comparing TD and OC Learning

Following Algorithm 1, we first contrast the parameter convergence of TD learning and OC learning with a toy simulation in which only the mean of stock returns are different in the two market regimes and all other market parameters are set equal. We consider a one-year investment horizon ($T = 1$) with 10 equal-step-size time points throughout the year for portfolio rebalancing ($\Delta t = 1/10$). The investor starts with $x_0 = \$1$ and sets a target of $z = \$1.4$ to be achieved by the end of the year. During the investment period, the investor explores the investment strategies with the exploration weight equal to $\xi = 0.5$. To mimic the stock market, we consider it has two regimes, “good” ($i = 1$) and “bad” ($i = 2$), which are alternating over time on a Markov Chain with generator $Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$. The “grounding true” mean and volatility of the market are $(\mu_{true,1}, \sigma_{true,1}) = (0.2, 0.2)$ in the “good” state and $(\mu_{true,2}, \sigma_{true,2}) = (-0.1, 0.2)$ in the “bad” state. For simplicity, suppose

the interest rates are $(r_1, r_2) = (0, 0)$, hence the “grounding true” Sharpe ratios are

$$\begin{aligned}\rho_{true,1} &= \frac{\mu_1 - r_1}{\sigma_1} = \frac{0.2 - 0}{0.2} = 1 \\ \rho_{true,2} &= \frac{\mu_2 - r_2}{\sigma_2} = \frac{-0.1 - 0}{0.2} = -0.5.\end{aligned}$$

To parameterize the value functions, we initialize the market parameters at

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}) \equiv (\sigma_1^{(0)}, \sigma_2^{(0)}, \rho_1^{(0)}, \rho_2^{(0)}) = (0.1, 0.1, 0.8, -0.3).$$

Finally, to ensure numerical stability during the training process, we set constraints for the range of the volatility $\sigma_1, \sigma_2 \in [0.1, 1]$ and the range of the Sharpe ratios $\rho_1, \rho_2 \in [-2, 2]$.

We summarize the configuration of this pilot simulation in Table 2.1. We set the learning rates to steadily decrease from their initial values to 1×10^{-5} over the training epochs.

Hyperparameters	Value at Initialization
T	1
Δt	0.1
Q	$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$
ξ	0.5
x_0	1
z	1.4
η for TD loss	$(1 \times 10^4, 1 \times 10^4, 2 \times 10^4, 2 \times 10^4)$
η for OC loss	$(1 \times 10^4, 1 \times 10^4, 1 \times 10^3, 1 \times 10^3)$
(r_1, r_2)	$(0, 0)$
θ_{true}	$(\sigma_{true,1} = 0.2, \sigma_{true,2} = 0.2, \rho_{true,1} = 1.0, \rho_{true,2} = -0.5)$
$\theta^{(0)}$	$(0.1, 0.1, 0.8, -0.3)$

Table 2.1: Configuration of the pilot simulation.

The paths of the market parameters over training epochs using the TD loss are given in Figure 2.1, which indicates that the market parameters did not converge to the “grounding true” values. While the Sharpe ratio parameters (ρ_1, ρ_2) converged to a wrong level, the volatility parameters (σ_1, σ_2) diverged, reaching the upper boundaries of the range we set previously. We also present the learning error of each parameter, defined by $\left| \frac{\theta_{learned} - \theta_{true}}{\theta_{true}} \right| \times$

100%, where $\theta_{learned}$ is the parameter we learned and θ_{true} is the corresponding true value. The poor convergence performance of TD learning is as expected. Because minimizing the TD loss is equivalent to minimizing the quadratic variation of M^θ (2.38), which should not be minimized.

Figure 2.2 shows the market parameters' updating paths for OC learning. All market parameters converged to their corresponding “grounding true” levels. While we acknowledge that the initial values for the Sharpe ratio parameters $(\rho_1^{(0)}, \rho_2^{(0)}) = (0.8, -0.3)$ are set close to their “grounding truths” $(\rho_{true,1}, \rho_{true,2}) = (1, -0.5)$, we reinitialize them as $(\rho_1^{(0)}, \rho_2^{(0)}) = (0.2, 0.2)$. Training with OC loss again produces converging paths of the market parameters towards the “grounding true” level, as shown in Figure 2.3. Our results empirically justify our solutions to the EMVRS problem in Theorem 2.3.1, while simultaneously revealing the superiority of OC learning over TD learning in this problem setting.

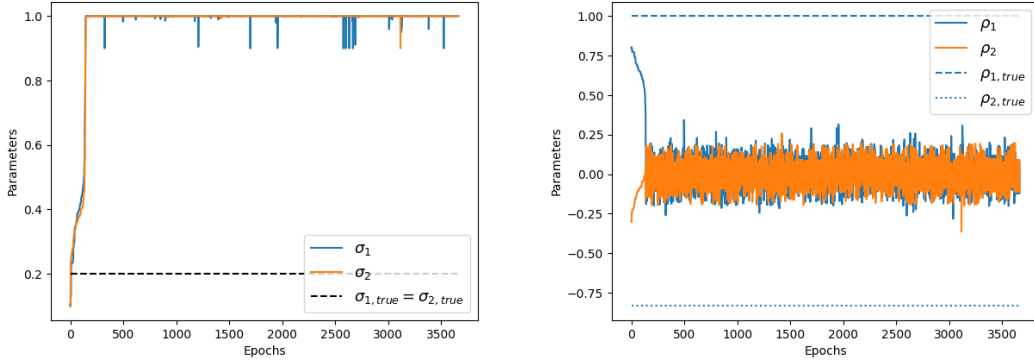


Figure 2.1: Parameter convergence of Temporal Difference (TD) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.8, -0.3)$. Learning errors of each parameter are (400%, 400%, 112.11%, 77.92%).

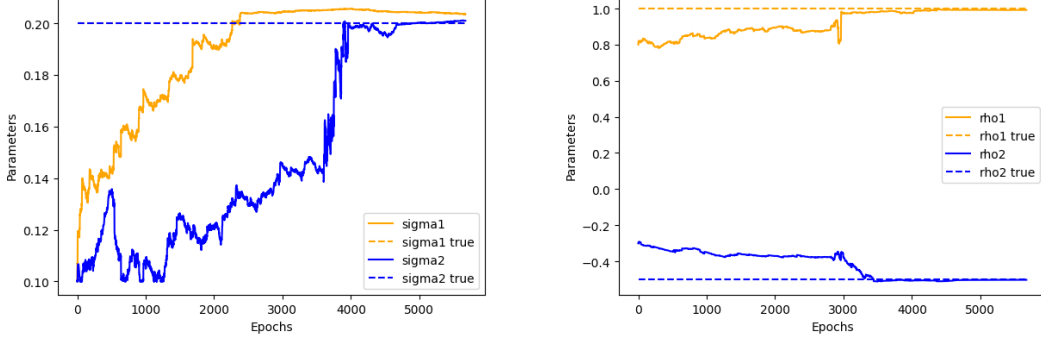


Figure 2.2: Parameter convergence of Orthogonality Condition (OC) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.8, -0.3)$. Learning errors of each parameter are (1.75%, 0.5%, 0.79%, 0.56%).

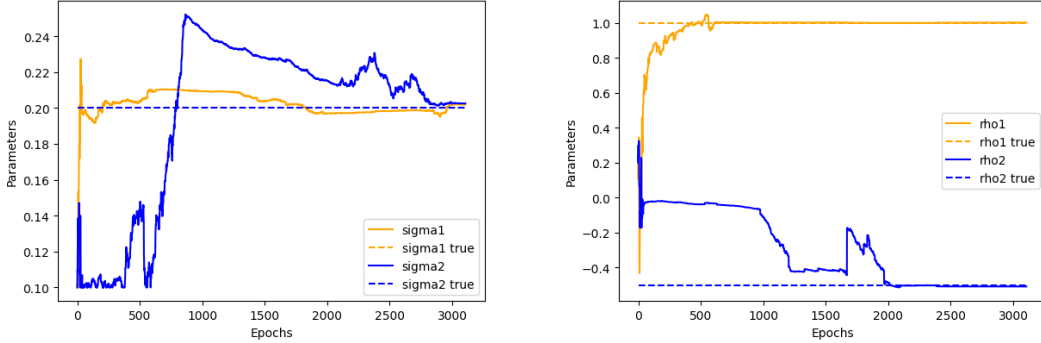


Figure 2.3: Parameter convergence using Orthogonality Condition (OC) loss. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.2, 0.2)$. Learning errors of each parameter are (0.95%, 1.25%, 0.14%, 0.18%).

Finally, we examine the robustness of OC learning in a more comprehensive simulation design, where the market regimes are more different from each other. Table 2.2 summarizes the configuration of this simulation study. We highlight the differences from the previous simulation: we considered the “grounding true” mean and volatility of the “good” state market to be $(\mu_{true,1}, \sigma_{true,1}) = (0.2, 0.2)$, as the US equity mean return and volatility were recorded at (18%, 18.4%) in 1996-2000 according to (Bae et al., 2014). For the “bad” state, we suppose $(\mu_{true,2}, \sigma_{true,2}) = (-0.2, 0.3)$, as the US equity mean return and volatility were

recorded at $(-17.4\%, 30.1\%)$ in 2006-2008 according to (Bae et al., 2014). Moreover, we assume the interest rate is lower in the “good” state and higher in the “bad” state, hence setting $(r_1, r_2) = (0.01, 0.05)$. So, the “grounding true” Sharpe ratios are

$$\begin{aligned}\rho_{true,1} &= \frac{\mu_1 - r_1}{\sigma_1} = \frac{0.2 - 0.01}{0.2} = 0.95, \\ \rho_{true,2} &= \frac{\mu_2 - r_2}{\sigma_2} = \frac{-0.2 - 0.05}{0.3} = -0.83.\end{aligned}$$

We set the market parameters at initialization as

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}) \equiv (\sigma_1^{(0)}, \sigma_2^{(0)}, \rho_1^{(0)}, \rho_2^{(0)}) = (0.1, 0.1, 0.5, -0.5)$$

Hyperparameters	Value at Initialization
T	1
Δt	0.1
Q	$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$
ξ	0.5
x_0	1
z	1.4
η for OC loss	$(1 \times 10^4, 1 \times 10^4, 1 \times 10^4, 1 \times 10^4)$
(r_1, r_2)	$(0.01, 0.05)$
θ_{true}	$(\sigma_{true,1} = 0.2, \sigma_{true,2} = 0.3, \rho_{true,1} = 0.95, \rho_{true,2} = -0.833)$
$\theta^{(0)}$	$(0.1, 0.1, 0.5, -0.5)$

Table 2.2: Configuration of the more comprehensive simulation.

The convergence of the market parameters is illustrated in Figure 2.4, which shows that all market parameters converged to a reasonably close range of their corresponding “grounding truths.” We emphasize that even though the market parameters of each regime are diverse in more dimensions, including the stock return volatility and interest rate, the time to convergence wasn’t significantly increased compared to that of the toy simulation.

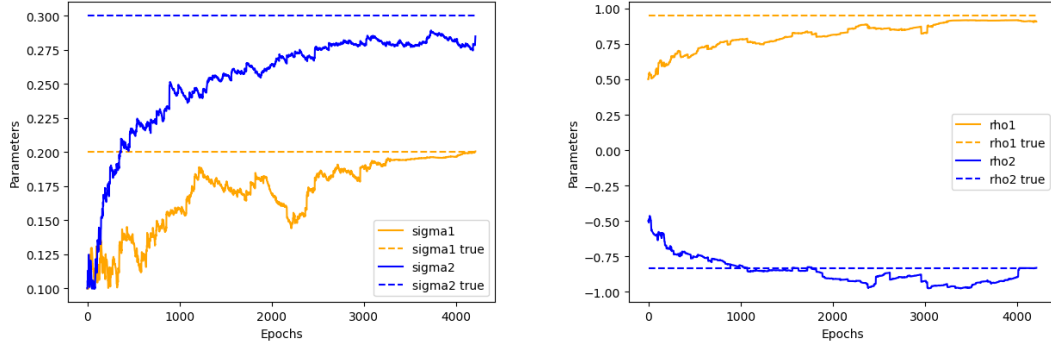


Figure 2.4: Parameter convergence of Orthogonality Condition (OC) learning. Market parameters are initialized at $\theta^{(0)} = (0.1, 0.1, 0.5, -0.5)$. Learning errors of each parameter are (0.25%, 5.07%, 4.45%, 0.34%).

2.5.2 Performance on Real Market Data: Comparing EMVRS and EMV

Having witnessed the parameters' convergence in simulation studies, we evaluate the investment performance of EMVRS on real market data, by comparing the mean and volatility of the terminal portfolio values. For simplicity, we consider an investor managing a market portfolio and financing under the risk-free interest rate. Hence, we collect from Yahoo Finance daily data of the S&P500 (ticker: “^GSPC”) market index as the market portfolio, and benchmark the 3-month US Treasury Bill (3mTbill) rate (ticker: “^IRX”) as the risk-free interest rate. We consider monthly portfolio rebalancing and set a 10-year rolling window starting from January 1st, 2006, with a one-month step size. So, the first window covers from January 1st, 2006 to December 31st, 2015, the second window covers from February 1st, 2006 to January 31st, 2016, and so on. We keep rolling for two years, hence the last window covers from December 1st, 2007 to November 30th, 2017, resulting in 24 10-year windows. Since the rolling windows cover the subprime mortgage crisis and the recovery since then, we broadly categorize the market into two regimes — a bullish regime and a bearish regime. On each rolling window, we apply the Viterbi algorithm (Viterbi, 1967) to identify the market regimes, then train the model using the parameters from the last time frame as the initial points. Figure 2.5 shows the regime-labeled time series of the first rolling window, where state 0 stands for bullish and 1 stands for bearish.

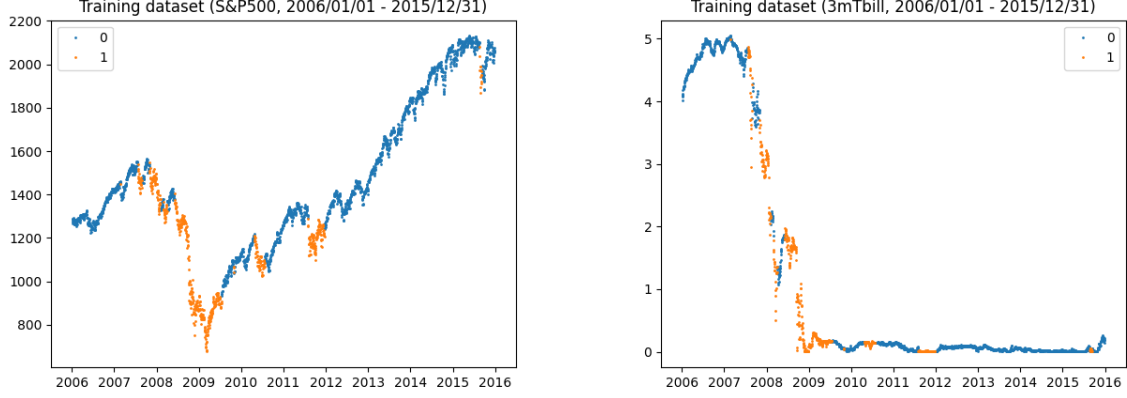


Figure 2.5: Regime-identified financial time series for training (left: S&P500, right: 3mT-bill). State 0 stands for bullish and state 1 stands for bearish.

Due to monthly rebalancing, we mark 120 time points on each 10-year window. We suppose the investor starts with an initial wealth $x_0 = \$1$ and targets the terminal portfolio value at $z = \$1.4$ by the end of ten years, which equates to a target annual return of approximately 3.422%. In order to maintain numerical stability throughout training, we standardize the S&P500 index and set an *action constraint* at 3, i.e., after sampling an action from the investment policy, we restrict the action value within $[-3x_0, 3x_0] = [-3, 3]$.⁴ We summarize the training configuration and provide the initial parameters for the first rolling window in Table 2.3.

⁴This in practice means the investor's leverage ratio is at most 3. The investor can borrow no more than 3 times of their initial wealth x_0 to invest in the stock index, or short sell no greater than $3x_0$ worth of stock index.

Hyperparameters	Value at Initialization
T	10
Δt	$\frac{1}{12}$
ξ	0.5
x_0	1
z	1.4
action constraint	3
η for TD loss	$(1 \times 10^3, 1 \times 10^3, 1 \times 10^3, 1 \times 10^3)$
η for OC loss	$(1 \times 10^3, 1 \times 10^3, 1 \times 10^3, 1 \times 10^3)$
$\theta^{(0)}$	$(0.2, 0.2, 1.0, 1.0)$

Table 2.3: Configuration of the real data study.

For comparison, we simultaneously train the EMVRS and EMV via both OC learning and TD learning and backtest the investment performance on the 24 10-year rolling windows.

We examine the trained models on various investment settings, including action constraints equal to 1, 1.5, 2, 3, and whether short selling is allowed. For example, an action constraint equal to 1 with short selling means the action can take values in $[-x_0, x_0] = [-1, 1]$, whereas action constraint equals to 1 without short selling means the action can take values in $[0, x_0] = [0, 1]$. We remind the readers that the actions are defined as the money values invested in the market portfolio (i.e., S&P500), hence an action taking a negative value is a short position. Under each investment setting, we independently trade 5 times over each of the 24 10-year rolling windows, resulting in 100 portfolio value trajectories. We then compute the mean and standard deviation of the last entry of these 100 trajectories as the estimates of the mean and volatility of the terminal portfolio values, respectively, which allows us to compute the annualized portfolio returns and volatilities. Since the risk-free interest rate is not constant in our problem, we computed the average terminal portfolio value if the investor holds the risk-free asset over the 24 10-year rolling windows, which is around \$1.0747. This yields the annualized risk-free rate at around 0.723%. Thereby, we can compute the Sharpe Ratios and we summarize the investment performances on the 24 10-year rolling windows in Table 2.4.

Under OC learning, EMVRS outperforms EMV in all of the investment settings we considered. Although we set a low investment target of portfolio return at 3.442%, the mean returns of the EMVRS all significantly exceed the predetermined target. Yet, the mean returns of EMV are fairly close to the target. EMV even fails to reach the target

mean return under one investment setting, when action constraint is 1 and short selling is not allowed. While both the mean and volatility of the portfolio returns are generally higher for EMVRS than those for EMV, the Sharpe Ratios of EMVRS are significantly higher than those of the EMV.

Under TD learning, EMVRS achieves lower mean portfolio returns but significantly inflated volatilities, which yields the lowest Sharpe Ratios under all investment settings. The performance of EMV is comparable to itself if trained with OC learning, with mean portfolio returns closed to the target. The Sharpe Ratios of EMVRS are not as good as those under OC learning. The poor performance aligns with our previous criticism of TD learning in our problem context.

Moreover, we also notice that under the same short selling setting, both the mean and volatility of the portfolio returns increase as the action constraint is gradually relaxed, resulting in gradually decreasing Sharpe Ratios. Except for EMVRS with OC learning, the performance on all other model settings demonstrates higher Sharpe Ratios when short selling is forbidden. These empirical observations remind us to be cautious when using high leverage and short positions in our investments.

2.6 Conclusion

Inspired by the past works on the regime-switching MV problem and the recent RL advances in stochastic control, we extend the classical MV portfolio optimization problem and formulate the EMVRS problem. We recognize that the stochastic control approach to the MV problems is optimal given that the market parameters are fully known. However, in practice the market parameters cannot be directly observed, which drives us to explore within the control space. Here, the RL framework offers an informed guidance of exploration within the control space, which is a Gaussian exploration with entropy regularization. Therefore, a combination of the RL framework and the stochastic control solution results in a collaborative exploration and exploitation of the optimal control policy. In this work, we derive an analytical solution to the EMVRS problem, which fills the gap between the regime-switching application in the stochastic control literature and the RL framework for the non-regime-switching MV problem.

Furthermore, we fully leverage the optimal stochastic control solution to the EMVRS problem, by adopting the functional form of the induced optimal value function and reparameterizing the value function with the market parameters. This differs from the approach in (H. Wang & Zhou, 2020), which reparameterizes the value function with a new set of

Training	Investment AC	Setting SS	EMVRS		EMV	
			Mean	Volatility	Mean	Volatility
OC Learning	1	✓	12.177%	1.932%	3.507%	1.673%
	1	✗	12.177%	1.932%	3.690%	0.857%
	1.5	✓	15.512%	2.890%	3.331%	2.419%
	1.5	✗	15.512%	2.890%	3.927%	1.246%
	2	✓	18.156%	3.849%	3.762%	3.226%
	2	✗	18.156%	3.849%	3.877%	1.736%
	3	✓	22.244%	5.771%	3.624%	4.917%
	3	✗	22.244%	5.771%	4.066%	2.176%
TD Learning	1	✓	7.319%	18.439%	3.426%	1.512%
	1	✗	8.694%	14.322%	3.690%	0.844%
	1.5	✓	8.342%	21.519%	3.406%	2.511%
	1.5	✗	10.002%	17.320%	3.937%	1.246%
	2	✓	8.759%	23.309%	3.590%	3.365%
	2	✗	10.720%	18.167%	3.887%	1.727%
	3	✓	9.095%	25.080%	3.117%	4.345%
	3	✗	11.394%	19.223%	4.081%	2.188%

Table 2.4: Performance on the 24 10-year windows. Mean and Volatility stand for the annualized mean and volatility of portfolio returns on the 24 10-year windows. AC is action constraint, SS is short selling, and SR is Sharpe Ratio. The target annual portfolio return is given by 3.422%.

parameters. Incorporating the induced optimal value function provides an informed and theory-backed starting point for the RL training algorithm. Using the market parameters as the parameterization of the RL algorithm reduces the dimension of the parameter space and produces more meaningful outputs when the model parameters have converged.

By adopting OC learning, our RL algorithm enables EMVRS to successfully recover the hidden market parameters in our simulation studies. The improved investment performance of EMVRS on the real market data also supports our proposed methods in the practical domain.

Chapter 3

Exploratory Mean-Variance with Jumps: An Equilibrium Approach

3.1 Introduction

The Mean-Variance (MV) portfolio optimization problem, first introduced by (Markowitz, 1952), has been the subject of extensive research. In the continuous-time setting, an MV investor aims to optimize her financial outcome by pursuing the highest possible portfolio return with the lowest risk, measured by portfolio volatility, at the end of the investment horizon. Jorion (1992) investigated the practical challenges of implementing the MV investment policy in real-world markets. Zhou and Li (2000) studied the Lagrangian dual of the MV problem as a stochastic linear-quadratic optimal control problem. Later on, Chiu and Li (2006) solved for the MV efficient frontier, and Xie et al. (2008) derived the optimal investment policy in an incomplete market setting; see (Zhang et al., 2018; Kalayci et al., 2019) for a broader overview of the past studies of the MV problem.

In this chapter, we take the stochastic control approach and extend the classical MV problem in two directions. In the first direction, we employ a Reinforcement Learning (RL) facilitated *exploratory formulation* of the MV problem. In the recent stochastic control literature, there is an emerging trend of applying RL techniques to classical stochastic control problems (H. Wang & Zhou, 2020; H. Wang et al., 2020; R. Jiang et al., 2022; Jia & Zhou, 2022a, 2022b, 2023; Dai et al., 2023; B. Wu & Li, 2024; Chen et al., 2025b). The fundamental technique is to replace the optimal solution to the classical stochastic control problem with a probability distribution (centered around the original optimal solution), in light of the Stochastic RL Algorithm by (Gullapalli, 1990). Such a distribution-valued

control is often called an *exploratory control* or a *policy* in the RL context. In this way, the exploratory control is capable of informed exploration within the control space, while maintaining greedy exploitation towards optimality. We point out here that the major motivation for such an exploratory formulation is the lack of knowledge of the controlled dynamics. Taking portfolio optimization as an example, the controlled dynamics is the portfolio value process that depends on the market dynamics. However, the precise law of the market evolution is unknown, and even a simple diffusion model requires knowledge of the drift and diffusion terms.

H. Wang et al. (2020) first introduced this “exploratory extension” to their controlled dynamics and solved for the optimal exploratory control of a linear-quadratic problem. They also proved the asymptotic equivalence between the classical solution and the exploratory solution. Later, Jia and Zhou (2022a, 2022b, 2023) leveraged the martingale properties of the optimal solutions to the exploratory-extended stochastic control problems and developed effective learning algorithms for RL policy evaluation and policy updates. H. Wang and Zhou (2020) studied the *Exploratory Mean-Variance* (EMV) problem, which solves for a distribution-valued investment policy centered on the classical MV solution. Since the EMV problem uses the Lagrangian dual of the classical MV problem, it is a time-consistent control problem and the proposed investment policy in (H. Wang & Zhou, 2020) is globally optimal at the beginning of the investment horizon.

In contrast, Dai et al. (2023) investigated the EMV problem as a time-inconsistent control (TIC) problem, as the time-inconsistency arises from the variance of terminal portfolio values, a quadratic function of the expected terminal portfolio value. Taking the TIC approach in (Björk et al., 2021), Dai et al. (2023) solved the Extended Hamiltonian-Jacobi-Bellman (EHJB) equations for an equilibrium policy, which is a probability distribution centered around the classical TIC solution provided in (Björk et al., 2021). While Dai et al. (2023) incorporated an exogenous and hidden state to model the unobservable market regimes or economic conditions, B. Wu and Li (2024) and Chen et al. (2025b) modeled the market regimes more directly with a time-homogeneous Markov chain and solved the EMV problem with regime-switching. Moreover, Chen et al. (2025b) also pointed out the significance of model parameter convergence as an evaluation criterion for RL models.

Our second extension to the classical MV problem is the incorporation of a jump component in the controlled dynamics, which is used to account for sudden shocks in the market. Jump-diffusion models have been widely applied to optimal stopping (Pham, 1997; Mordecki, 2002; Bayraktar & Egami, 2008), option evaluation (Amin, 1993; Andersen & Andreasen, 2000; Carr & Mayo, 2007; Clift & Forsyth, 2008; Toivanen, 2008; Xu et al., 2009), and hedging problems (Park et al., 1993; He et al., 2006; Mina et al., 2015). The MV problem with jump-diffusion controlled dynamics is studied by (Øksendal

& Sulem, 2005) in the classical stochastic control formulation, and by (Bender & Thuan, 2023) in the exploratory control formulation. More recently, Gao et al. (2024) introduced the exploratory formulation of stochastic control problems with jump-diffusion controlled dynamics. However, all of them considered the Lagrangian dual of the MV problem, hence following the time-consistent control approach.

Differing from (Øksendal & Sulem, 2005; Bender & Thuan, 2023; Gao et al., 2024), this chapter tackles the EMV problem with a jump-diffusion by considering it as a TIC problem and solving for an equilibrium solution, an approach adopted by (Björk et al., 2021; Dai et al., 2023). We establish the exploratory formulation of the MV problem with jump-diffusion market dynamics, which we call the *Exploratory Mean-Variance with Jumps (EMVJ) problem*. We derive the EHJB equations and solve for an equilibrium investment policy in explicit form. Our solution aligns with (Björk et al., 2021) if we remove the exploration and the jumps in the dynamics. Following (Jia & Zhou, 2022a), we then utilize the martingale properties of our analytical solutions and develop an Actor-Critic RL algorithm. Through iterative interactions with the market, our RL model is able to gradually understand the market behavior and learn the equilibrium investment policy, without the knowledge of the underlying market dynamics. Our numerical results demonstrate the effectiveness of our RL algorithm on simulated data, as well as the practical applicability of our RL model on real market data.

We make two main contributions to the literature. First, we fill a gap in the literature by solving the EMV problem with jump-diffusion dynamics via the TIC approach. We note that Gao et al. (2024) solved the Lagrangian dual of the MV problem, so did many other existing works; see (H. Wang & Zhou, 2020; Bender & Thuan, 2023; B. Wu & Li, 2024; Chen et al., 2025b). Their investment strategies are optimal at the beginning of an investment horizon subject to a prespecified investment target, disregarding the time-inconsistent preferences and actions an investor can have. Our equilibrium investment policy is a game-theoretical result, supposing an investor competes with her future incarnations throughout the investment horizon. Therefore, our equilibrium investment policy is the best action an investor can take, at any given time during the investment horizon. Second, we fully leverage our analytical solutions in the design of a RL model, by parameterizing the RL model with market parameters that have practical meanings. Our choice of parameterization not only enables us to check whether the parameters have converged to the true values in our simulation study, but also significantly simplifies the RL model¹ without sacrificing its effectiveness.

The remainder of this chapter is structured as follows. We review the necessary back-

¹As described in Section 3.4, our RL model has only 3 parameters.

ground and preliminaries in Section 3.2. In Section 3.3, we introduce the EMVJ problem and present the verification theorem and analytical solutions. We propose a RL algorithm in Section 3.4 and present our numerical results in Section 3.5. Finally, we conclude in Section 3.6.

3.2 Problem Formulation and Preliminaries

We adopt the notation from (Øksendal & Sulem, 2005). Consider a finite decision horizon T and a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$. Let $L := \{L_t\}_{t \in [0, T]}$ be a Lévy process and $\Delta L_s := L_s - L_{s-}$ be the jump of L at time s . Denote by $\mathbf{B}_0$ the collection of Borel sets of $\mathbb{R}$ whose closure does not contain 0. Then, for $B \in \mathbf{B}_0$, define the Poisson random measure of L as

$$N(t, B) = \sum_{s: 0 \leq s \leq t} 1_B(\Delta L_s), \quad (3.1)$$

and the Lévy measure of L as

$$\nu(B) = \mathbb{E}[N(1, B)]. \quad (3.2)$$

Intuitively, $N(t, B)$ represents the number of jumps of sizes in a Borel set B up to time t , whereas $\nu(B)$ measures the expected number of jumps of sizes in B per unit time. For any $B \in \mathbf{B}_0$, $\{N(t, B) : t \in [0, T]\}$ is a Poisson process with intensity $\nu(B)$.

Define the compensated Poisson random measure as

$$\tilde{N}(dt, dz) = N(dt, dz) - \nu(dz)dt. \quad (3.3)$$

We know from (Øksendal & Sulem, 2005) that for any $B \in \mathbf{B}_0$, $\{\tilde{N}(t, B) : t \geq 0\}$ is a martingale. The Lévy measure is $\nu(dz) = \zeta_J F_J(dz)$, where $\zeta_J > 0$ is the arrival rate of Poisson jumps per unit time and $F_J(z)$ is the probability distribution of jump size z . Here, we require that the Lévy measure should satisfy

$$\int_{\mathbb{R}} \min(1, z^2) \nu(dz) < \infty \quad \text{and} \quad \int_{\mathbb{R}} e^{2z} \nu(dz) < \infty. \quad (3.4)$$

For simplicity, we consider investing in a market with one risky asset (stock) and one risk-free asset. The risk-free rate is a constant $r \geq 0$. The stock price is driven by

$$\frac{dS_t}{S_{t-}} = \mu dt + \sigma dW_t + \int_{\mathbb{R}} (e^z - 1) \tilde{N}(dt, dz), \quad (3.5)$$

where $\mu \in \mathbb{R}, \sigma > 0$ are constant market parameters and $\{W_t\}_{t \in [0, T]}$ is a one-dimensional Brownian Motion on the filtered probability space. Under the second condition of (3.4), we have that $\mathbb{E}[S_t]$ and $\mathbb{E}[S_t^2]$ are finite for any $t \in [0, T]$, according to (Cont & Tankov, 2003).

An investor controls the money (in dollar value) that she invests in the stock, reallocating the portfolio in response to the behavior of the market. The discounted self-financing portfolio value $X := \{X_t\}_{t \in [0, T]}$ follows the process

$$dX_t^u = u_t(\mu - r)dt + u_t\sigma dW_t + \int_{\mathbb{R}} u_t(e^z - 1)\tilde{N}(dt, dz), \quad (3.6)$$

where $u := \{u_t := u(t, X_t^u)\}_{t \in [0, T]}$ denotes the control, i.e., money invested in the stock.

Remark 3.2.1 *Note that in reality the market parameters, (μ, σ) , and the jump parameters, including ζ_J and the parameters of $F_J(z)$ (the jump size distribution), are unknown. So, investing under a control u , as a deterministic function of time t and wealth x that depends on the market parameters and the jump parameters, is impractical. While one possible solution is to estimate the unknown parameters from the market data, for example using Maximum Likelihood Estimation, this method often suffers from estimation error as the observed market data is just one realization of the market dynamics; see (Luenberger, 2013) for the mean-blur problem as a typical example. Alternatively, we can adopt the idea of Stochastic Reinforcement Learning (Gullapalli, 1990) that employs a random investor exploring for the optimal investment policy, within the control space. In this way, we avoid estimation error as there is no parameter estimation involved. Instead, the parameters are learned while the investor interacts with the dynamic market. In the recent stochastic control literature, such an investor randomization is achieved by extending the classical control u to a probability distribution valued policy, π , often centered at u ; see (H. Wang & Zhou, 2020; H. Wang et al., 2020; Jia & Zhou, 2022b; R. Jiang et al., 2022; Jia & Zhou, 2023; Gao et al., 2024; Chen et al., 2025b) for examples of such extensions.*

Furthermore, we recognize the difficulty of investing under a control u in practice, as the market parameters, (μ, σ) , and the jump parameters, including ζ_J and the parameters of $F_J(z)$ (the jump size distribution), are unknown. Henceforth, we consider an *exploratory control*, $\pi := \{\pi_{t, X_t^\pi}(u) := \pi(u; t, X_t^\pi)\}_{t \in [0, T]}$, that enables exploration within the control space. Here, $\pi : (t, x) \in [0, T] \times \mathbb{R} \mapsto \pi_{t, x} \in \mathcal{P}(\mathbb{R})$ is a stochastic feedback policy, which maps a time and portfolio value pair (t, x) to a probability density function on the control space $\mathbb{R}$. For each $t \in [0, T]$, let $u_t^\pi := u(t, X_t^\pi)$ be the investment control sampled from

$\pi_{t,X_t^\pi}(u)$. Then, a sampled portfolio value process is given by

$$dX_t^\pi = u_t^\pi(\mu - r)dt + u_t^\pi \sigma dW_t + \int_{\mathbb{R}} u_t^\pi(e^z - 1)\tilde{N}(dt, dz). \quad (3.7)$$

To account for the stochasticity of the randomly sampled control, we consider an enlargement of the filtered probability space. Following (Jia & Zhou, 2022b, 2023), we suppose that the probability space is rich enough to support mutually independent copies of a random variable, $\{U_t, t \in [0, T]\}$, each uniformly distributed on $[0, 1]$. Denote by $(\Omega, \mathcal{F}, \{\mathcal{G}_t\}_{t \in [0, T]}, \bar{\mathbb{P}})$ the new filtered probability space. Here, $\mathcal{G}_s = \mathcal{F} \vee \sigma(U_t, 0 \leq t \leq s)$ is the σ -algebra generated by $\mathcal{F}_s$ and the uniform random variables up to time s , and $\bar{\mathbb{P}}$ is a product extension from $\mathbb{P}$, which is now defined on the extended σ -algebra $\mathcal{G}_T$. Following the arguments in (Gao et al., 2024), we also introduce an extended Poisson random measure, $N'(dt, dz, dv)$, on the product space $[0, T] \times \mathbb{R} \times [0, 1]$, where v is a realization of a Uniform random vector $V \sim \text{UNIF}([0, 1])$. The intensity measure of N' is given by $dt \nu(dz)dv$ and its compensator is $\tilde{N}'(dt, dz, dv) := N'(dt, dz, dv) - dt \nu(dz)dv$. Furthermore, given a distribution d , denote $G^d : [0, 1] \rightarrow \mathbb{R}$ by the quantile function of d . Then, for some (t, x) , a sampled action u from the policy distribution $\pi_{t,x}$ can be realized by first sampling $v \sim \text{UNIF}([0, 1])$ and then applying the quantile function, i.e., $u = G^{\pi_{t,x}}(v)$.

Therefore, as we randomly sample trajectories of a control u^π from a policy π and take the average of the portfolio values across trajectories, the sampled portfolio value process in (3.7) follows the same distribution as

$$dX_t^\pi = \int_{\mathbb{R}} u(\mu - r)\pi_{t,X_t^\pi}(u)du dt + \sqrt{\int_{\mathbb{R}} u^2 \sigma^2 \pi_{t,X_t^\pi}(u)du} dW_t + \int_{\mathbb{R} \times [0, 1]} G^{\pi_{t,x}}(v)(e^z - 1)\tilde{N}'(dt, dz, dv), \quad (3.8)$$

which we refer to as the *exploratory portfolio value process*.

Definition 3.2.1 (Admissible Policy) *We say that $\pi := \{\pi_{t,x}(\cdot)\}_{t,x \in [0, T] \times \mathbb{R}_+}$ is an admissible policy if the following conditions are satisfied:*

- (1) For any $(t, x) \in [0, T] \times \mathbb{R}$, $\pi_{t,x}(u) \in \mathcal{P}(\mathbb{R})$;
- (2) $\int_{\mathbb{R}} u^2 \pi_{t,x}(u)du \leq C_1(1 + x^2)$, for any $(t, x) \in [0, T] \times \mathbb{R}$;
- (3) $\left| \int_{\mathbb{R}} u(\pi_{t,x}(u) - \pi_{t,y}(u))du \right|^2 + \left| \sqrt{\int_{\mathbb{R}} u^2 \pi_{t,x}(u)du} - \sqrt{\int_{\mathbb{R}} u^2 \pi_{t,y}(u)du} \right|^2 + \int_{[0, 1]} |G^{\pi_{t,x}}(v) - G^{\pi_{t,y}}(v)|^2 dv \leq C_2|x - y|^2$, for any $t \in [0, T], x, y \in \mathbb{R}$;

(4) $\int_{\mathbb{R}} \pi_{t,x}(u) |\log \pi_{t,x}(u)| du < C_3(1 + |x|^2)$, for any $(t, x) \in [0, T] \times \mathbb{R}$.

Here, C_1, C_2, C_3 are constants that are independent of t and x . The set of all admissible policies is denoted by $\mathcal{A}$.

We conclude this section with the following lemma, which connects the definition of an admissible policy to the solution of the SDE (3.8). The proof is given in Appendix A.2.1

Lemma 3.2.0.1 *Following an admissible policy $\pi \in \mathcal{A}$, the exploratory portfolio value process (3.8) guarantees a unique càdlàg adapted solution X^π such that*

$$\mathbb{E} \left[\sup_{t \in [0, T]} (X_t^\pi)^2 \right] < \infty.$$

3.3 Exploratory Mean-Variance with Jumps Problem

The MV portfolio optimization problem aims to find an investment policy that achieves the highest return and the lowest portfolio volatility. Since the investment control u is randomly sampled from the exploratory control π , we consider entropy regularization to encourage exploration within the control space. Denote by $H(\pi) := - \int_{\mathbb{R}} \pi(u) \log \pi(u) du$ the entropy of a distribution-valued exploratory policy π . The objective function is given by

$$\begin{aligned} J(t, x; \pi) &= \mathbb{E}_{t,x} \left[X_T^\pi + \lambda \int_t^T H(\pi_s) ds \right] - \frac{\gamma}{2} \text{Var}_{t,x}(X_T^\pi) \\ &= \mathbb{E}_{t,x} \left[X_T^\pi - \frac{\gamma}{2} (X_T^\pi)^2 + \lambda \int_t^T H(\pi_s) ds \right] + \frac{\gamma}{2} (\mathbb{E}_{t,x}[X_T^\pi])^2, \end{aligned} \quad (3.9)$$

where $\lambda > 0$ is the exploration weight and $\gamma > 0$ is the risk aversion factor. Additionally, $\mathbb{E}_{t,x}(\cdot)$, $\text{Var}_{t,x}(\cdot)$ respectively denote the expectation and variance conditioned on $X_t^\pi = x$, i.e., $\mathbb{E}_{t,x}(\cdot) \equiv \mathbb{E}(\cdot | X_t^\pi = x)$ and $\text{Var}_{t,x}(\cdot) \equiv \text{Var}(\cdot | X_t^\pi = x)$. This means that the expectation is taken under the condition that the portfolio value process X_t^π starts from x at time t , following a given policy π . Hence, the Exploratory Mean-Variance with Jumps (EMVJ) portfolio optimization problem is

$$\sup_{\pi \in \mathcal{A}} J(t, x; \pi). \quad (3.10)$$

We point out here that following an admissible policy $\pi \in \mathcal{A}$ as mentioned in Definition 3.2.1 also ensures $J(t, x; \pi) < \infty$.

Remark 3.3.1 We note that the entropy regularization term in the objective function (3.9) plays a critical role in two major aspects. First, maximizing the classical MV objective $\mathbb{E}_{t,x}[X_T^\pi] - \frac{\gamma}{2}\text{Var}_{t,x}(X_T^\pi)$ with an additional entropy term enforces the exploratory policy π to avoid collapsing too quickly to a single deterministic control. This balances exploitation (picking known highly rewarding controls) with exploration (keeping the policy spread out), enabling possible discovery of outperforming controls that used to be unknown to the investor. More practically, the inclusion of an exploration weight λ also allows us to control the extent to which an investor explores in the control space. Second, the entropy term acts as a convex regularizer and yields optimal policies that are Boltzmann-like (H. Wang et al., 2020). It not only smooths the optimization landscape but also produces stochastic policy distributions (often Gaussian in Linear-Quadratic settings such as under our problem setup) that are easier to learn and evaluate.

The inclusion of the variance term² in the objective function (3.9) introduces time-inconsistency, because it implies a nonlinear dependence on the conditional distribution of future wealth, violating Bellman’s principle of optimality (Björk et al., 2021). When the investor re-optimizes at a later time with updated wealth and conditional expectation, her objective changes because the square of a conditional expectation introduces nonseparability over time. This leads to preference reversal, meaning that an investor who initially commits to an optimal strategy would deviate from it later in time once the optimization is recomputed. To address this, we can replace the classical notion of optimality with a *game-theoretic equilibrium control*, also known as a *time-inconsistent equilibrium control*, obtained by local (rather than global) optimality with respect to unilateral deviations in infinitesimal time. It is called an “equilibrium” because the investment strategy is characterized as a subgame perfect Nash equilibrium of an intra-personal game between the investor’s present and future selves. In this sense, under the equilibrium control, each future self has no incentive to deviate, though it is generally not globally optimal. We next introduce a formal definition of an *equilibrium investment policy*.

Definition 3.3.1 Let $\pi^* \in \mathcal{A}$ be an admissible investment policy. For any $(t, x) \in [0, T] \times \mathbb{R}$, consider a perturbed policy $\pi^\varepsilon := \{\pi_{s,y}^\varepsilon(\cdot)\}_{(s,y) \in [t,T] \times \mathbb{R}}$, defined as

$$\pi_{s,y}^\varepsilon(u) = \begin{cases} v(u), & (s, y) \in [t, t + \varepsilon) \times \mathbb{R}, \\ \pi_{s,y}^*(u), & (s, y) \in [t + \varepsilon, T] \times \mathbb{R}, \end{cases}$$

²Recall that $\text{Var}_{t,x}(X) = \mathbb{E}_{t,x}(X^2) - (\mathbb{E}_{t,x}(X))^2$.

where $\varepsilon > 0$, and $v(\cdot) \in \mathcal{P}(\mathbb{R})$ is an arbitrary density function satisfying

$$\int_{\mathbb{R}} u^2 v(u) du < \infty \quad \text{and} \quad \int_{\mathbb{R}} v(u) |\log v(u)| du < \infty,$$

which implies $\pi^\varepsilon \in \mathcal{A}$. Suppose that for any $(t, x) \in [0, T] \times \mathbb{R}_+$ and any such perturbed strategy π^ε , we have

$$\limsup_{\varepsilon \downarrow 0} \frac{J(t, x; \pi^\varepsilon) - J(t, x; \pi^*)}{\varepsilon} \leq 0.$$

Then, π^* is an equilibrium investment policy for problem (3.10), and $J(t, x; \pi^*)$ is the corresponding value function.

Define the *infinitesimal generator* for $\psi(t, x) \in C^{1,2}([0, T] \times \mathbb{R})$ with $\sup_x \{\psi_{xx}(t, x)\} < \infty$ and $\pi \in \mathcal{A}$ as

$$\begin{aligned} \mathcal{L}^\pi \psi(t, x) &= \int_{\mathbb{R}} \left[u(\mu - r)\psi_x + \frac{1}{2}\sigma^2 u^2 \psi_{xx} \right] \pi_{t,x}(u) du \\ &\quad + \int_{\mathbb{R} \times [0,1]} [\psi(t, x + G^{\pi_{t,x}}(v)(e^z - 1)) - \psi(t, x) - G^{\pi_{t,x}}(v)(e^z - 1)\psi_x] \nu(dz) dv \end{aligned} \quad (3.11)$$

$$\begin{aligned} &= \int_{\mathbb{R}} \left[u(\mu - r)\psi_x + \frac{1}{2}\sigma^2 u^2 \psi_{xx} \right. \\ &\quad \left. + \int_{\mathbb{R}} [\psi(t, x + u(e^z - 1)) - \psi(t, x) - u(e^z - 1)\psi_x] \nu(dz) \right] \pi_{t,x}(u) du, \end{aligned} \quad (3.12)$$

where the last equality is given by the construction of $u = G^{\pi_{t,x}}(v)$ and thereby the fact that

$$\int_{\mathbb{R} \times [0,1]} f(t, x, G^{\pi_{t,x}}(v), z) \nu(dz) dv = \int_{\mathbb{R}} \int_{\mathbb{R}} f(t, x, u, z) \nu(dz) \pi_{t,x}(u) du,$$

for some function f .

The next two theorems establish the theoretical core of this chapter, which are critical foundations to design the RL algorithms in the subsequent section.

Theorem 3.3.1 (Verification Theorem) *Suppose that there exist two functions $V(t, x)$ and $g(t, x) \in C^{1,2}([0, T] \times \mathbb{R})$ such that the following conditions are satisfied.*

(i) For any $(t, x) \in [0, T] \times \mathbb{R}$,

$$\sup_{\pi \in \mathcal{A}} \left\{ (\partial_t + \mathcal{L}^\pi)V + \lambda H(\pi) + \gamma g^\pi \mathcal{L}^\pi g^\pi - \frac{\gamma}{2} \mathcal{L}^\pi (g^\pi)^2 \right\} = 0. \quad (3.13)$$

(ii) Let $\pi_{t,x}^*(u)$ be the maximizer that achieves the supremum in (3.13), and suppose it satisfies, for any $(t, x) \in [0, T] \times \mathbb{R}$,

$$(\partial_t + \mathcal{L}^\pi)g^\pi = 0. \quad (3.14)$$

(iii) For any $x \in \mathbb{R}$ and $\pi \in \mathcal{A}$,

$$V(T, x) = g(T, x; \pi) = x. \quad (3.15)$$

Then $\pi^* := \{\pi_{t,x}^*(\cdot)\}_{(t,x) \in [0,T] \times \mathbb{R}}$ is an equilibrium policy for problem (3.10). Moreover, we have the corresponding value function

$$V(t, x) = J(t, x; \pi^*), \quad (3.16)$$

and the auxiliary value function

$$g^{\pi^*}(t, x) \equiv g(t, x; \pi^*) := \mathbb{E}_{t,x}(X_T^{\pi^*}). \quad (3.17)$$

Theorem 3.3.2 (Equilibrium Policy and Value Functions) *An equilibrium investment policy for problem (3.10) is given by*

$$\pi_{t,x}^*(u) \sim N\left(\frac{\mu - r}{\gamma(\sigma^2 + \delta^2)}, \frac{\lambda}{\gamma(\sigma^2 + \delta^2)}\right), (t, x) \in [0, T] \times \mathbb{R} \quad (3.18)$$

where $\delta^2 = \int_{\mathbb{R}} (e^z - 1)^2 \nu(dz)$.

The corresponding value function is $V(t, x) = x + C(t)$ and the corresponding auxiliary value function is $g(t, x; \pi^*) = x + h^*(t)$, where

$$C(t) = (T - t) \left[\frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right], \quad (3.19)$$

$$h^*(t) = (T - t) \left[\frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} \right]. \quad (3.20)$$

The proofs of Theorems 3.3.1 and 3.3.2 are given in Appendices A.2.2 and A.2.3 respectively.

3.4 RL Algorithm

In this section, we design an RL algorithm to iteratively learn the equilibrium investment policy without the knowledge of the market and jump dynamics. We leverage the martingale property, in particular the *Orthogonality Condition*, from the analytical results in Theorems 3.3.1 and 3.3.2 to define a loss function, which we use to optimize our RL model.

Recall that the stock price dynamics is

$$\frac{dS_t}{S_{t-}} = \mu_{true} dt + \sigma_{true} dW_t + \int_{\mathbb{R}} (e^z - 1) \tilde{N}(dt, dz), \quad (3.21)$$

where $(\mu_{true} \in \mathbb{R}, \sigma_{true} > 0)$ are the true mean and volatility of stock returns. We consider Merton's jump-diffusion model, which was studied by (Gao et al., 2024); see also (Merton, 1976; Bates, 1991; Das, 2002). The jump density is

$$\nu(dz) = \zeta_{J,true} \frac{1}{\sqrt{2\pi\sigma_{J,true}^2}} \exp\left(-\frac{(z - \mu_{J,true})^2}{2\sigma_{J,true}^2}\right) dz, \quad (3.22)$$

where $\zeta_{J,true} > 0$ is the true arrival rate of jumps per unit time, and $(\mu_{J,true} \in \mathbb{R}, \sigma_{J,true} > 0)$ are the true mean and standard deviation of jump sizes. Most importantly, they are unknown parameters.

Moreover, we note from (3.18) that the equilibrium investment policy depends on μ, σ , and another parameter δ , with

$$\begin{aligned} \delta_{true}^2 &:= \int_{\mathbb{R}} (e^z - 1)^2 \nu(dz) \\ &= \zeta_{J,true} \left[\exp(2\mu_{J,true} + 2\sigma_{J,true}^2) - 2\exp(\mu_{J,true} + \frac{1}{2}\sigma_{J,true}^2) + 1 \right]. \end{aligned} \quad (3.23)$$

We aggregate the jump parameters $(\zeta_J, \mu_J, \sigma_J)$ into δ , and define the parameterization $\theta := (\mu, \sigma, \delta)$. The parameterized policy is $\pi^\theta := \{\pi_{t,x}^\theta(\cdot)\}_{(t,x) \in [0,T] \times \mathbb{R}}$, with

$$\pi_{t,x}^\theta(u) \sim N\left(\frac{\mu - r}{\gamma(\sigma^2 + \delta^2)}, \frac{\lambda}{\gamma(\sigma^2 + \delta^2)}\right), (t, x) \in [0, T] \times \mathbb{R}. \quad (3.24)$$

Under π^θ , the portfolio value process in (3.8) becomes

$$\begin{aligned} dX_t^\theta &= \frac{(\mu_{true} - r)(\mu - r)}{\gamma(\sigma^2 + \delta^2)} dt + \sqrt{\sigma_{true}^2 \left(\frac{\lambda}{\gamma(\sigma^2 + \delta^2)} + \frac{(\mu - r)^2}{\gamma^2(\sigma^2 + \delta^2)^2} \right)} dW_t \\ &\quad + \left(\frac{\mu - r}{\gamma(\sigma^2 + \delta^2)} \right) \int_{\mathbb{R}} (e^z - 1) \tilde{N}(dt, dz). \end{aligned} \quad (3.25)$$

Inspired by the HJB (3.13), we define, for a parameterized policy $\pi^\theta \in \mathcal{A}$, a stochastic process $M^\theta := \{M_t^\theta := M(t, X_t^\theta; \pi^\theta)\}_{(t, X_t^\theta) \in [0, T] \times \mathbb{R}}$ such that $\forall (t, x) \in [0, T] \times \mathbb{R}$,

$$M_t^\theta = V^\theta(t, X_t^\theta) - \frac{\gamma}{2}(g^\theta(t, X_t^\theta))^2 + \lambda \int_0^t H(\pi_s^\theta) ds, \quad (3.26)$$

where $V^\theta(t, x) = x + C^\theta(t)$ and $g^\theta(t, x) = x + h^\theta(t)$ are the value function and the auxiliary value function corresponding to the parameterized policy π^θ , respectively. Here, C^θ and h^θ are respectively given by (3.19) and (3.20), with π replaced by π^θ .

Theorem 3.4.1 *The process M^θ is an $(\{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$ martingale when $\theta = \theta_{true} := (\mu_{true}, \sigma_{true}, \delta_{true})$.*

The proof of this theorem is postponed to Appendix A.2.4.

Additionally, we know that any square-integrable martingale $M := \{M_t\}_{t \in [0, T]}$ has an orthogonality condition

$$\mathbb{E} \left[\int_0^T \xi_t dM_t \right] = 0,$$

where the test function $\xi := \{\xi_t\}_{t \in [0, T]}$ is an arbitrary $\{\mathcal{F}_t\}_{t \in [0, T]}$ -adapted process that is square-integrable with respect to M . This, together with the martingale property of M^θ , allows us to define an *Orthogonality Condition (OC) Loss* function. Following (Jia & Zhou, 2022a), we choose the test function as the partial derivative of the parameterized value function V^θ with respect to its parameters θ^3 .

Definition 3.4.1 (Orthogonality Condition (OC) Loss) *Take $\theta = (\mu, \sigma, \delta) = (\theta_1, \theta_2, \theta_3)$. For $(t, x) \in [0, T] \times \mathbb{R}$ and $j = 1, 2, 3$, the OC loss is given by*

$$L_{OC}(\theta_j) = \mathbb{E}_{t, x} \left[\int_0^T \frac{\partial V^\theta}{\partial \theta_j}(t, X_t^{\pi^\theta}) dM_t^\theta \right], \quad (3.27)$$

with

$$\frac{\partial V^\theta}{\partial \theta_1}(t, x) = \frac{\partial V^\theta}{\partial \mu} = (T - t) \left[\frac{(\mu - r)}{\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right], \quad (3.28)$$

$$\frac{\partial V^\theta}{\partial \theta_2}(t, x) = \frac{\partial V^\theta}{\partial \sigma} = -(T - t) \left[\frac{(\mu - r)^2 \sigma}{\gamma(\sigma^2 + \delta^2)^2} + \frac{\lambda \sigma}{(\sigma^2 + \delta^2)} \right], \quad (3.29)$$

$$\frac{\partial V^\theta}{\partial \theta_3}(t, x) = \frac{\partial V^\theta}{\partial \delta} = -(T - t) \left[\frac{(\mu - r)^2 \delta}{\gamma(\sigma^2 + \delta^2)^2} + \frac{\lambda \delta}{(\sigma^2 + \delta^2)} \right]. \quad (3.30)$$

³In addition to staying aligned with (Jia & Zhou, 2022a), this test function is also inspired by the θ updating rule if we instead use a gradient descent method to optimize the value function V^θ .

Remark 3.4.1 *The orthogonality condition loss, first introduced by (Jia & Zhou, 2022a), is built on the fact that under the true value function, the temporal difference residual (obtained from Itô's formula and the HJB equation) must be a martingale with zero drift, hence orthogonal to every square-integrable test function. Intuitively, if V^θ correctly approximates the true value function, i.e., $\theta = \theta_{\text{true}}$, then no predictable component should remain in the temporal difference residual, resulting in zero conditional expectation. Note that under our RL setting, the parameterized policy π^θ and the corresponding value function V^θ share the same parameters, thus the policy optimization and the value function approximation are achieved simultaneously when $\theta = \theta_{\text{true}}$. Leveraging the equivalence of parameter convergence and the martingale property from Theorem 3.4.1, the orthogonality condition loss enforces effective updates of policy and value functions, by iteratively updating θ and reducing the loss to zero.*

Considering a discretization $0 = t_0 < \dots < t_N = T$ with mesh size Δt , the discretized loss function becomes

$$L_{OC}(\theta_j) = \sum_{n=0}^{N-1} \frac{\partial V^\theta}{\partial \theta_j}(t_n, X_{t_n}^{\pi^\theta}) \left\{ V^\theta(t_{n+1}, X_{t_{n+1}}^{\pi^\theta}) - V^\theta(t_n, X_{t_n}^{\pi^\theta}) - \frac{\gamma}{2} \left[g^\theta(t_{n+1}, X_{t_{n+1}}^{\pi^\theta})^2 - g^\theta(t_n, X_{t_n}^{\pi^\theta})^2 \right] + \frac{\lambda}{2} \log \left(\frac{2\pi e \lambda}{\gamma(\sigma^2 + \delta^2)} \right) \Delta t \right\}. \quad (3.31)$$

We summarize the whole training procedure in Algorithm 2.

Algorithm 2: Training with Orthogonality Condition Loss

Initialize the number of epochs N_{epochs} , the investment horizon T , the mesh size of the continuous time discretization Δt , the exploration parameter $\lambda > 0$, the initial portfolio value $x_0 > 0$, the learning rates $\eta = (\eta_1, \eta_2, \eta_3)$, the “grounding true” market parameters $\theta_{true} = (\theta_{true,1}, \theta_{true,2}, \theta_{true,3}) \equiv (\mu_{true}, \sigma_{true}, \delta_{true})$, and the model parameters $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}) = (\mu^{(0)}, \sigma^{(0)}, \delta^{(0)})$.

Compute $N = \frac{T}{\Delta t}$.

for $k = 0, \dots, N_{epoch} - 1$ **do**

 Fix a realized path of the Brownian motion $\{W_{t_n}^{(k)}\}_{n=0}^N$, with

$\Delta W_{t_n}^{(k)} := W_{t_{n+1}}^{(k)} - W_{t_n}^{(k)} \sim N(0, \Delta t)$.

 Sample number of jumps $N_k \sim Poisson(\zeta_{J,true} \Delta t)$ and sample the jumps

$(Z_1^{(k)}, \dots, Z_{N_k}^{(k)}) \stackrel{iid}{\sim} N(\mu_{J,true}, \sigma_{J,true}^2)$.

 Simulate $\{X_{t_n}^{(k)}\}_{n=0}^N$ via (3.25), with (μ, σ, δ) replaced by $(\mu^{(k)}, \sigma^{(k)}, \delta^{(k)})$,
 $\{W_{t_n}\}_{n=0}^N$ replaced by $\{W_{t_n}^{(k)}\}_{n=0}^N$, and the integral replaced by $\sum_{j=1}^{N_k} (e^{Z_j^{(k)}} - 1)$.

 Compute $\{\frac{\partial V^{\theta^{(k)}}}{\partial \mu}(t_n, X_{t_n}^{(k)})\}_{n=0}^N$, $\{\frac{\partial V^{\theta^{(k)}}}{\partial \sigma}(t_n, X_{t_n}^{(k)})\}_{n=0}^N$ and $\{\frac{\partial V^{\theta^{(k)}}}{\partial \delta}(t_n, X_{t_n}^{(k)})\}_{n=0}^N$
 via (3.28), (3.29) and (3.30), with (μ, σ, δ) replaced by $(\mu^{(k)}, \sigma^{(k)}, \delta^{(k)})$.

for $j = 1, 2, 3$ **do**

 Compute $L(\theta_j^{(k)}) := L_{oc}(\theta_j^{(k)}; \{X_{t_n}^{(k)}\}_{n=0}^N)$ via (3.31).

 Update $\theta_j^{(k+1)} \leftarrow \theta_j^{(k)} + \eta_j \times L(\theta_j^{(k)})$

end

end

3.5 Numerical Results

In this section, we train and evaluate our model on simulated data and real market data. On simulated data (Subsection 3.5.1), we train our model and show that the model parameters converge to the corresponding true values. Moreover, the investment performance (in terms of mean terminal portfolio value) on the simulated data is close to the theoretical expectation. We also examine our model on real market data (Subsection 3.5.2), which illustrates the potential profitability of our equilibrium policy.

3.5.1 Analysis On Simulated Market Data

From Yahoo Finance, we collect the S&P500 market index (total return) as a “representative” stock at a daily frequency, ranging from January 1st of 2000 to December 31st of 2023. We fit Merton’s jump-diffusion model, parameterized by both the market parameters (μ, σ) and the jump parameters $(\mu_J, \sigma_J, \zeta_J)$. Denote by $\theta := (\mu, \sigma, \mu_J, \sigma_J, \zeta_J)$. Consider a time discretization $0 = t_0 < \dots < t_N = T$ with mesh size Δt . When Δt is small, we suppose the log-returns $R_i := \log\left(\frac{S_{t_i}}{S_{t_{i-1}}}\right)$, for $i = 1, \dots, \frac{T}{\Delta t}$, are approximately independently and identically distributed with density function

$$f(r; \theta) = \sum_{m=0}^{\infty} p_m(\zeta_J, \Delta t) \cdot \phi(r; \nu_m(\theta, \Delta t), \tau_m^2(\theta, \Delta t)),$$

where $p_m(\zeta_J, \Delta t) = \frac{e^{-\zeta_J \Delta t} (\zeta_J \Delta t)^m}{m!}$ is the Poisson probability of m jumps, $\phi(\cdot; \nu, \tau^2)$ is the Gaussian density with mean ν and variance τ^2 , and

$$\begin{aligned} \nu_m(\theta, \Delta t) &= \left(\mu - \frac{1}{2} \sigma^2 - \lambda(e^{\mu_J + \frac{1}{2} \sigma_J^2} - 1) \right) \Delta t + m \mu_J \\ \tau_m^2(\theta, \Delta t) &= \sigma^2 \Delta t + m \sigma_J^2. \end{aligned}$$

While the direct computation of an infinite sum $f(r, \theta)$ is impractical, we approximate it with a truncated sum up to $m=2$. Because multiple jumps per interval are very rare and $p_m(\zeta_J, \Delta t)$ for $m > 2$ are negligible. We calculate the likelihood function as $L(\theta; R) = \prod_{i=1}^{T/\Delta t} f(R_i; \theta)$ and apply the Nelder-Mead method (Nelder & Mead, 1965) to obtain the maximum likelihood estimates (MLE) of θ . We choose the MLEs of θ as the “grounding true” parameter values $(\mu_{true}, \sigma_{true}, \zeta_{J,true}, \mu_{J,true}, \sigma_{J,true})$ for simulation. In this way, our simulated data reflects the real market.

We consider a one-year investment horizon ($T = 1$) with daily rebalancing ($\Delta t = \frac{1}{252}$), reflecting an assumption of 252 trading days per year. The investor starts with an initial wealth of $x_0 = 1$, and invests in the stock following the equilibrium policy (3.18) with initial parametrization $(\mu_0, \sigma_0, \delta_0) = (0.1, 0.1, 0.05)$. For simplicity, we suppose there is no cost of borrowing, i.e., the risk-free rate $r = 0$. The exploration weight λ and the risk aversion parameter γ are set as 1 for training. We summarize all the parameters in Table 3.1.

Parameters	Values
x_0	1
$(T, \Delta t)$	$(1, \frac{1}{252})$
λ	1
γ	1
r	0.01
$(\mu_{true}, \sigma_{true})$	(0.0878, 0.1321)
$\zeta_{J,true}$	27.6813
$(\mu_{J,true}, \sigma_{J,true})$	(-0.0040, 0.0274)
δ_{true} (by (3.23))	0.1449
δ_0	0.05
(μ_0, σ_0)	(0.1, 0.1)

Table 3.1: Parameters for training and evaluation on simulated data.

We train the model following Algorithm 2 and all the model parameters converge to their “grounding true” values in 550 epochs as shown in Figure 3.1. This indicates that our model has been optimized to the equilibrium policy. We make a note here that the three parameters do not converge at the same rate — in the case showed in Figure 3.1, μ converges fastest while σ converges slowest.⁴ In the training process, we impose a linear learning rate scheduler to control the learning rates for each parameter.

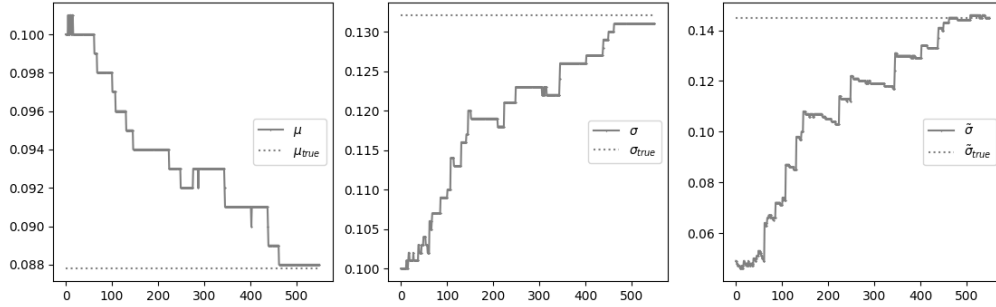


Figure 3.1: Convergence of μ (left), σ (middle) and δ (right).

⁴The speed of parameter convergence depends on how close the initial points are to the corresponding true values.

Next, we check if our trained model is able to achieve the theoretically expected investment performance. From Theorem 3.3.2, we know that if an investor follows a policy π^* , the expected terminal portfolio value is

$$\begin{aligned}\mathbb{E}_{0,x_0}(X_T^{\pi^*}) &= g(t, x; \pi^*) = x_0 + h^{\pi^*}(t) \\ &= x_0 + (T - t) \left[\frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} \right].\end{aligned}$$

We suppose the investor follows the mean of the equilibrium policy (3.18), that is, at each time t , she invests $u_t = \frac{\mu}{\gamma(\sigma^2 + \delta^2)}$ dollars in the stock, and saves the remaining $X_t - u_t$ in cash at the risk-free interest rate r . If $u_t > X_t$, this means the investor is investing with leverage, and the cost of borrowing is equal to the risk-free interest rate r . If $u_t < 0$, this means the investor is short-selling the stock, and saving the premium at the risk-free interest rate r . We highlight that investing at our policy mean is an extension to the equilibrium policy considered in (Björk et al., 2021), as they coincide with each other when there are no jumps in the market dynamics. While we train the model with risk aversion parameter $\gamma = 1$, we examine the investment performance for investors with different risk appetites, $\gamma \in [0.1, 0.5, 1, 2, 5]$. Higher γ values represent more risk averse investors, hence holding a smaller long or short position in the stock. For each investor, we simulate 100 stock price trajectories and make investments repeatedly. As a result, we get 100 portfolio value trajectories. The *realized mean* and *realized volatility* of the terminal portfolio value are calculated as the average and the standard deviation of the last entry over the 100 portfolio value trajectories. We also compute and contrast the *empirical objective function*, $J_{\text{empirical}}$ using (3.9) with $\lambda = 0$,⁵ and the theoretical value of the objective function V given the equilibrium policy, using the value function in Theorem 3.3.2.

We summarize the investment performance in Table 3.1. For the realized mean, realized volatility and $J_{\text{empirical}}$, we also compute the Monte Carlo standard errors following the metrics in (Morris et al., 2019), which are reported in the parentheses in Table 3.1. The realized mean terminal portfolio values are pretty close to the corresponding theoretical means. The realized volatility diminishes as the investor becomes more risk averse (i.e., as γ increases), which is practically meaningful as more aggressive investment results in higher volatility. The empirical objective function values are reasonably close to the theoretical values, and the gap narrows as the investor becomes more risk averse.

⁵Because we are following the mean of the equilibrium policy, there is no policy exploration.

γ	realized mean	theoretical mean	realized volatility	$J_{empirical}$	theoretical V
0.1	2.9980 (0.3032)	3.0213	3.0316 (0.2154)	2.5384 (0.3055)	2.0106
0.5	1.3996 (0.0606)	1.4043	0.6063 (0.0431)	1.3077 (0.0611)	1.2021
1	1.1998 (0.0303)	1.2021	0.3032 (0.0215)	1.1538 (0.0305)	1.1011
2	1.0999 (0.0152)	1.1011	0.1516 (0.0108)	1.0770 (0.0153)	1.0505
5	1.0400 (0.0061)	1.0404	0.0606 (0.0043)	1.0308 (0.0061)	1.0202

Table 3.2: Investment performance on simulated market data. The Monte Carlo standard errors of the realized means, realized volatilities and $J_{empirical}$ are reported in parentheses.

3.5.2 Analysis on Real Market Data

To evaluate our model on real market data, we choose the S&P500 market index mentioned in the last subsection as the risky asset. We also collect the US 3-month Treasury Bill (3mTBill) from Yahoo Finance as the risk-free interest rate. The frequency of the 3mTBill data is daily and the range is from January 1st, 2000 to December 31st, 2023, which matches the S&P500 data.

For model training, we divide the investment horizon of 24 years, from 2000 to 2023, into 14 eleven-year rolling windows, 2000-2010, 2001-2011, $\dots$, 2013-2023. Each rolling window spans from the first day of the starting year to the last day of the ending year, e.g., window 2000-2010 covers the period from January 1st, 2000 to December 31st, 2010. The first 10 years of each window is the *training period* and the final year is the *evaluation period*. We suppose an investor is risk averse with $\gamma = 5$. In the training phase, we set $\lambda = 5$ to encourage exploration within the control space. Over each rolling window, we fit a jump-diffusion model on the training period and use the MLEs as the initial parameters of the RL model. We evaluate the model performance on the evaluation period of each rolling window.

The investment performances on the training periods and the evaluation periods are given in Tables 3.3 and 3.4, respectively, including the terminal portfolio value means, volatilities and Sharpe Ratios (SRs). Over each training or evaluation period, we simultaneously create 100 self-financing portfolios each with initial value of $x_0 = 1$. We rebalance the portfolios on a daily basis according to the stock and risk-free interest rate dynamics, following the equilibrium policy (3.18). To reduce the portfolio volatility, we set $\lambda = 0.01$ if we invest over the training periods and set $\lambda = 0.1$ if we invest over the evaluation periods. The *mean* and the *volatility* columns are calculated as the average and standard deviation of the 100 portfolios' values on the last day of the investment period. The Monte Carlo

standard errors are reported in the parentheses, following the metrics provided in (Morris et al., 2019). The *risk-free* column values are the terminal values if we hold the risk-free asset throughout the investment period.

window	mean	volatility	risk-free
2000-2010	1.3980 (0.0198)	0.1978 (0.0141)	1.3071
2001-2011	1.3859 (0.0202)	0.2022 (0.0144)	1.2348
2002-2012	1.3845 (0.0185)	0.1848 (0.0131)	1.1944
2003-2013	1.8165 (0.0201)	0.2007 (0.0143)	1.1766
2004-2014	1.8445 (0.0187)	0.1869 (0.0133)	1.1654
2005-2015	1.9569 (0.0140)	0.1397 (0.0099)	1.1499
2006-2016	1.5993 (0.0150)	0.1500 (0.0107)	1.1148
2007-2017	1.5213 (0.0154)	0.1535 (0.0109)	1.0667
2008-2018	1.6454 (0.0146)	0.1463 (0.0104)	1.0311
2009-2019	3.6082 (0.0207)	0.2066 (0.0147)	1.0370
2010-2020	4.9670 (0.0195)	0.1945 (0.0138)	1.0570
2011-2021	5.5411 (0.0272)	0.2720 (0.0193)	1.0592
2012-2022	7.4664 (0.0231)	0.2311 (0.0164)	1.0591
2013-2023	5.2763 (0.0225)	0.2246 (0.0160)	1.0795

Table 3.3: Investment performance on training periods. The model was trained with $(\lambda, \gamma) = (5, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.01, 5)$. The Monte Carlo standard errors of the portfolio value means and volatilities are reported in parentheses.

Clearly, Table 3.3 shows that all the portfolios invested over the training periods achieve higher mean returns than the risk-free return, with reasonably low volatility. The mean terminal portfolio values are very high during the post subprime crisis in 2008 — the portfolio value triples in 2009-2019 and even grows sevenfold in 2012-2022, in spite of the breakout of COVID since 2020. Even in the training periods that are close to or cover the subprime crisis, the training performance is still reliable, as the mean terminal portfolio value grows at least by around 50% in 2004-2014, 2005-2015, 2006-2016, 2007-2017 and 2008-2018. This implies that our models are well-trained and the market parameters of the training periods are properly learned through the training process.

Table 3.4 conveys the generalizability and practical applicability of our model. In 12 out of the 14 evaluation periods, the mean terminal portfolio values are higher than the risk-free portfolio, with reasonably low volatility. The only two exceptions are in 2018 and 2022, during which the market trend flips, as shown in Figure 3.2. That is, the market

trend of the training period (2008-2017 and 2012-2021) is opposite to that of the evaluation period (2018 and 2022, respectively). So, the model is misled by the training period. For example, in the window 2012-2022, the model is trained over a bullish period of 2012-2021 and it will assume the market to continue growing in 2022, which is not the case due to the breakout of COVID. As a result, the model decides to hold a long position in the stock during a bearish evaluation period in 2022. However, in the windows 2009-2019 and 2013-2023, our model takes the years of 2018 and 2022 as part of the training periods and adjusts the investment strategies, leading to an improved performance in the evaluation periods in 2019 and 2023. This implies that our model is robust against different market conditions, as long as the market does not significantly violate the model assumptions. Moreover, our model preserves the potential of reversing the adverse performance, by gradually digesting the market dynamics and adjusting the investment strategies accordingly.

For purposes of comparison, we also present the investment performance using the MLEs of a fitted Merton jump-diffusion model as the parameters of the equilibrium investment policy. We use the subscript *MLE* to denote the investment performances using MLEs in Table 3.4. Except for 2018 and 2022, our model achieves higher Sharpe Ratios in all other evaluation periods, compared to the investment strategy using MLEs. Since our model takes the MLEs as the initial parameters in the training phase, the improved investment performance demonstrates the effectiveness of our algorithm. In the years 2018 and 2022, the investments using MLEs also obtain negative Sharpe Ratios, which provides a poor starting point for us to train the model. This partially explains why our model performance in 2018 and 2022 is poor.

Finally, we also consider a more conservative investor with $\lambda = 1$. After retraining and retesting the model, we improve the investment performance in 2018, as shown in Table 3.5. This reiterates our model’s robustness against violation of model assumptions in the real-world market.

year	mean	volatility	SR
2018	1.0598 (0.0151)	0.1507 (0.0107)	0.2679

Table 3.5: Investment performance in 2018. The model was retrained with $(\lambda, \gamma) = (1, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.01, 5)$. SR denotes the Sharpe Ratio. The Monte Carlo standard errors of the portfolio value mean and volatility are reported in parentheses.

year	mean	volatility	SR	risk-free	$mean_{MLE}$	$volatility_{MLE}$	SR_{MLE}
2010	1.0666 (0.0173)	0.1726 (0.0123)	0.3784	1.0013	1.0189 (0.0121)	0.1214 (0.0086)	0.1452
2011	1.0227 (0.0186)	0.1863 (0.0132)	0.1197	1.0004	1.0067 (0.0129)	0.1292 (0.0092)	0.0487
2012	1.0621 (0.0105)	0.1054 (0.0075)	0.5820	1.0008	1.0226 (0.0081)	0.0813 (0.0058)	0.2677
2013	1.2297 (0.0101)	0.1006 (0.0071)	2.2791	1.0005	1.1121 (0.0077)	0.0774 (0.0055)	1.4414
2014	1.2132 (0.0124)	0.1244 (0.0088)	1.7127	1.0002	1.0481 (0.0077)	0.0773 (0.0055)	0.6187
2015	1.0372 (0.0170)	0.1696 (0.0121)	0.2171	1.0004	1.0095 (0.0108)	0.1083 (0.0077)	0.0837
2016	1.0625 (0.0099)	0.0985 (0.0070)	0.6029	1.0030	1.0457 (0.0089)	0.0894 (0.0064)	0.4773
2017	1.0880 (0.0043)	0.0427 (0.0030)	1.8451	1.0093	1.0752 (0.0040)	0.0404 (0.0029)	1.6309
2018	0.9975 (0.0141)	0.1412 (0.0100)	-0.1548	1.0194	1.0098 (0.0118)	0.1184 (0.0084)	-0.0814
2019	1.4672 (0.0138)	0.1384 (0.0098)	3.2274	1.0206	1.2435 (0.0107)	0.1069 (0.0076)	2.0840
2020	1.6298 (0.0476)	0.4761 (0.0338)	1.3158	1.0034	1.2601 (0.0345)	0.3449 (0.0245)	0.7445
2021	1.8431 (0.0173)	0.1726 (0.0123)	4.8834	1.0003	1.2414 (0.0108)	0.1075 (0.0076)	2.2434
2022	0.2339 (0.0356)	0.3555 (0.0252)	-2.2119	1.0202	0.7706 (0.0225)	0.2246 (0.0160)	-1.1109
2023	1.5263 (0.0201)	0.2063 (0.0147)	2.3002	1.0517	1.1563 (0.0124)	0.1239 (0.0088)	0.8440

Table 3.4: Investment performance on evaluation periods. The model was trained with $(\lambda, \gamma) = (5, 5)$ and the table is calculated with $(\lambda, \gamma) = (0.1, 5)$. SR denotes the Sharpe Ratio, and the subscript MLE denotes the investment performance calculated using MLEs of Merton's jump-diffusion model. The Monte Carlo standard errors of the portfolio value means and volatilities are reported in parentheses.

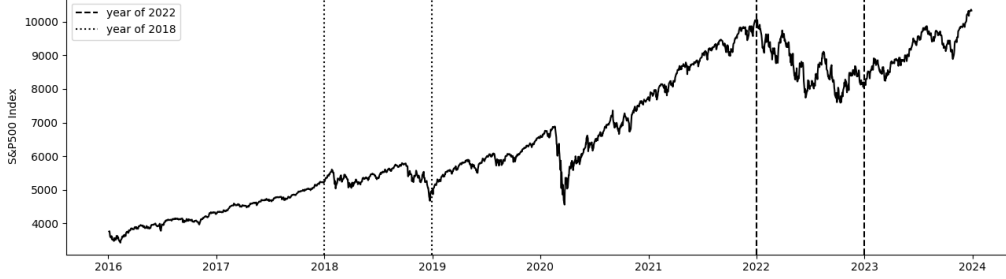


Figure 3.2: S&P500 market index. The period between the dotted vertical lines is the year 2018 and the period between the dashed vertical lines is the year 2022. The market was significantly shocked due to the crash of the [cryptocurrency bubble](#) in 2018, as well as the post-COVID historical [global interest rate hike](#) and the escalation in [the Russo-Ukrainian war](#) in 2022.

3.6 Conclusion

In this chapter, we apply the time-inconsistent control (TIC) approach to a RL-facilitated exploratory MV problem with jump-diffusion market dynamics. Unlike many past works which choose to solve the Lagrangian dual of the MV problem (H. Wang & Zhou, 2020; Gao et al., 2024; B. Wu & Li, 2024; Chen et al., 2025b), we tackle the time-inconsistency of the MV problem following the methods introduced in (Björk et al., 2021). We acknowledge that the uniqueness of an equilibrium solution to a TIC problem remains as an open research topic, as discussed in (Dai et al., 2023). Therefore, we disregard the uniqueness argument of our solution, and decide to present one solution by establishing and solving the EHJB equations. We present a verification theorem (Theorem 3.3.1) that ensures our solution is an equilibrium solution, according to Definition 3.3.1.

Furthermore, we choose the RL model parameterization that is practically meaningful, based on our analytical solutions in Theorem 3.3.2. We then leverage the martingale property, specifically the orthogonality condition, to define the Orthogonality Condition (OC) loss function, which is more effective in our problem context compared to another widely-applied Temporal-Difference (TD) loss in RL, as discussed in (Jia & Zhou, 2022a; Chen et al., 2025b). With our choice of parameterization and loss function, we are able to check whether parameters can converge to the corresponding true values. Thankfully, all parameters converge correctly in our simulation study, indicating that our RL algorithm is effective to learn the equilibrium policy for jump-diffusion market dynamics.

On real market data, we examine our model over a course of 24 years since 2000. Our model performs well on all the training periods and is profitable in 13 out of 14 evaluation periods, with reasonably low volatility. We acknowledge that our model can be misled if the market significantly violates the model assumptions. Yet, under a careful and conservative investment setting, our model can still make a profit. Moreover, our model preserves the capability of reversing the loss in successive investment periods, by continuously interacting and learning from the market.

Chapter 4

Exploratory Investment-Consumption Problem with Non-Exponential Discounting

4.1 Introduction

The investment-consumption problem in the stochastic control literature begins with Merton's seminal contributions (Merton, 1969, 1971), which establish the foundational framework to optimize consumption and portfolio allocation for an investor facing uncertain investment returns and income cashflow. Later, with several choices of utility functions, Karatzas et al. (1986, 1987a, 1987b) and Cox and Huang (1989) solved for the explicit solution via the Dynamic Programming Principle (DPP) approach and the martingale convex duality approach, with finite and infinite time horizons. In Merton's model, investors allocate wealth between a risk-free bond and a risky asset to maximize expected utility of consumption and terminal wealth. This paradigm assumes full observability of market parameters, such as the expected return (drift) and volatility of risky assets, and it has since inspired a rich literature in finance, economics, and statistics. In practice, however, financial market parameters are rarely known with certainty. For instance, the drift of risky assets is typically unobservable and must be inferred from observed price data. Yet, a simple Maximum Likelihood Estimation suffers from the *mean-blur problem* as documented by (Luenberger, 2013). More broadly, parameters of the market dynamics such as the volatility, correlation, and jump intensities are also subject to estimation error and cannot be directly observed. This motivates the study of portfolio optimization under

parameter uncertainty.

A well-established approach in this direction is *Bayesian learning*, which treats unobservable parameters as random variables with prior distributions that are dynamically updated through Bayes' rule as new market data becomes available (Brennan, 1998; Karatzas & Zhao, 2001; Cvitanić et al., 2006; Pham, 2011). This framework transforms the original partial-information control problem into a full-information one augmented by belief states, thereby yielding dynamically adaptive strategies that hedge against not only market risk but also parameter uncertainty. An alternative and complementary strand of research addresses parameter uncertainty using *ambiguity models*, which replaces a single prior with a family of plausible priors. Classical frameworks include the *maxmin expected utility* (Gilboa & Schmeidler, 1989), the α -*maxmin model* (Marinacci, 2002; Ghirardato et al., 2004; Maccheroni et al., 2006), and the *smooth ambiguity model* (Klibanoff et al., 2005). Unlike Bayesian learning, which reduces uncertainty through statistical updating, ambiguity models emphasize robustness to misspecification by penalizing deviations across priors.

Recent advances in Reinforcement Learning (RL) further tackle the parameter uncertainty. RL provides data-driven and exploratory methods to learn optimal policies in complex or partially observed environments without estimation of model parameters. Particularly, recent stochastic control literature has shown interest in RL techniques (Jia & Zhou, 2022a, 2022b, 2023) and achieved improvement compared to classical stochastic control solutions. The key motivation for applying RL to these problems lies in the dilemma that model parameters (such as the mean and volatility of the returns of risky assets) are unobservable in reality. Classical control solutions to these problems give the optimal investment strategy, provided that the parameters are known. When parameters are unknown, RL allows for a policy exploration within an admissible control space, often implemented by sampling actions from a probability distribution centered around the classical control solution. Over iterations of trial and error, the RL model gradually learns the optimal strategy. In financial applications, H. Wang and Zhou (2020) solved the mean-variance portfolio optimization with a precommitted policy, and B. Wu and Li (2024) and Chen et al. (2025b) extended the problem to a regime-switching market. R. Jiang et al. (2022) solved a portfolio optimization problem with the Kelly criterion. Moreover, Dai et al. (2023) revisited the mean-variance portfolio optimization problem and solve for the equilibrium solution with the Time Inconsistent Control (TIC) approach. This perspective complements the Bayesian and ambiguity approaches, offering a flexible methodology for sequential decision-making problems in finance.

Beyond uncertainty quantification, investor preferences introduce an additional layer of complexity. Intertemporal preferences can be decomposed into *state preferences*, reflecting

attitudes toward one’s wealth status and consumption level, and *time preferences*, capturing individual’s preference for consumption today against tomorrow, often referred to as the time value of money (Myers, 1968). The state preference has been extensively studied in the optimal investment-consumption literature by applying various types of utility functions. Kraft et al. (2013) employed a recursive utility to capture the investor’s intertemporal trade-offs and incorporated market incompleteness where risks cannot be fully hedged with traded assets. Constantinides (1990) introduced habit formation¹ to tackle the consumption smoothing puzzle inherent in Merton’s problem. This leads to a strand of research that incorporates habit formation into utility maximization; see (Detemple & Zapatero, 1992; Detemple & Karatzas, 2003; Munk, 2008; Yu, 2017; van Bilsen et al., 2020; Y. Wang et al., 2024). Moreover, Dai et al. (2023) studied the state preferences in the mean-variance portfolio optimization context.

The time preference, in contrast, plays a central role in shaping long-term behavior through the choice of discounting functions. While exponential discounting accounts for a risk-neutral preference, non-exponential discounting accounts for time-varying preferences and inevitably generates time-inconsistency. The optimal investment-consumption literature has shown a steady interest in the non-exponential discounting control problem; see (Strotz, 1955; Epstein & Hynes, 1983; Laibson, 1997; Barro, 1999) for some of the seminal works and also (Krusell & Smith, 2003; Zhao et al., 2014; Ebert et al., 2020; Y. Wang et al., 2024) for more recent contributions. Björk et al. (2021), along with their past works (Björk & Murgoci, 2010, 2014a, 2014b; Björk et al., 2016), introduced the fundamental framework of TIC problems and cover the non-exponential discounting control as a special case. TIC problems account for time-varying preferences by considering that an investor at a given time point competes with their incarnations at future time points, leading to a game-theoretical equilibrium solution. See also (Ekeland & Pirvu, 2008; Yong, 2011; Hu et al., 2012; Yong, 2012; Hu et al., 2017).

In this chapter, we focus on the time preference by studying the optimal investment-consumption problem under non-exponential discounting and handle parameter uncertainty with RL-facilitated policy exploration. Adopting the log utility as covered by (Björk et al., 2021), we formulate an exploratory version of the investment-consumption problem, named as the *Exploratory Investment-Consumption (EIC)*, using entropy regularization. Noticing that time-inconsistency arises from non-exponential discounting, we explicitly solve for an equilibrium² policy for investment and consumption, which can be expressed

¹Habit formation implies that past living standards affect the utility at any given consumption level.

²In time-inconsistent control, we use the term *equilibrium* instead of *optimal*, to reflect the game-theoretical nature of the solution. This is also partially because the uniqueness of an equilibrium solution is an open question (Dai et al., 2023).

as two independent probability distributions centered around their corresponding classical solutions. Specifically, we find out that the investment policy follows a Gaussian distribution and the consumption policy follows a Gamma distribution. Our result is practically meaningful because investment can take positive and negative values depending on the investor’s long and short positions in the asset, while consumption cannot be negative and the support of the Gamma distribution is $\mathbb{R}_+$. Beyond that, we analyze the change of an investor’s investment-consumption policies when facing uncertainty in the discounting rates, modeled by the non-exponential discounting. We show that the mean and variance of the Gamma consumption policy both decrease, and the variance of the Gaussian investment policy decreases. That is, comparing to exponential discounting, an investor using non-exponential discounting tends to take more conservative moves, shrinking and maintaining a low level of consumption as well as avoiding shuffling their portfolio significantly. Furthermore, we design a RL training algorithm to learn the equilibrium policy, parameterized by the market parameters. To ensure accurate convergence of the model parameters, we take the Policy Evaluation approach (Jia & Zhou, 2022a) and the q-learning approach (Jia & Zhou, 2023), respectively deriving two loss functions — the *Orthogonality Condition (OC) Loss* and the *q-learning loss*. As a result, our simulation study demonstrates that the model parameters converge to their corresponding true values without any parameter estimation, indicating that the RL model has learned the equilibrium investment and consumption policies. Finally, we highlight three original contributions of this chapter to the literature.

- (i) To the best of our knowledge, this is the first work that applies RL techniques to solve a non-exponential discounting dual-control problem, following the TIC equilibrium approach. Complementary to (Dai et al., 2023) that tackle time-inconsistency from the state preferences, we model time-inconsistency from the time preferences via non-exponential discounting functions and explicitly solve for an equilibrium policy for investment and consumption.
- (ii) We theoretically study the change in behaviors of investors using exponential and non-exponential discounting functions. According to (Ebert et al., 2020), we remark that investors using non-exponential discounting face uncertain discounting rates, and we show that under our exploratory formulation investors will behave more prudently in both consumption and investment. Our theoretical results cover the classical non-exploratory setting, where an investor with non-exponential discounting will only reduce consumption.
- (iii) We develop an implementable and tractable RL algorithm in two learning approaches, enabling iterative parameter updates while the RL model interacts with the market

dynamics. By parameterizing the RL model with the market parameters, our choice of RL model parameters have practical meanings and therefore the model convergence directly reflects the level at which the model perceives market dynamics. Moreover, we provide a Policy Iteration Theorem, which theoretically guarantees that the RL model will converge to the equilibrium policy.

The remainder of this chapter is structured as follows. Section 4.2 reviews the Investment-Consumption problem with log utility and non-exponential discounting (Björk et al., 2021). Then, we extend it to its exploratory version in Section 4.3 and solve for the analytical solution in Section 4.4, where we also provide a verification theorem and a policy iteration theorem. In Section 4.5, we derive the Orthogonality Condition loss and the q-learning loss and propose a RL training algorithm. We examine our algorithm in Section 4.6 via a simulation study, with a gamma discounting function and selected parameters. We conclude with Section 4.7.

4.2 Classical Optimal Investment-Consumption Problem

In this section, we review the optimal investment-consumption problem with non-exponential discounting and without reinforcement learning, which serves as a benchmark. Since non-exponential discounting gives rise to the issue of time inconsistency, we adopt the time-consistent approach of (Björk et al., 2021) to solve for equilibrium strategies. Via a formal verification argument, we present the analytical form of the corresponding equilibrium investment-consumption strategies.

Consider an investor who manages a portfolio and makes withdrawals for consumption until a finite time $T > 0$. For ease of presentation, the market is simplified to consist of one risky asset, the stock $\{S_t\}_{t \in [0, T]}$, and one risk-free asset, the bond $\{R_t\}_{t \in [0, T]}$. Denote by $\{W_t\}_{t \in [0, T]}$ a one-dimensional Brownian Motion defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = \{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$ that satisfies the usual conditions. The stock and the bond prices are respectively driven by

$$\begin{aligned} dS_t &= \mu S_t dt + \sigma S_t dW_t, \\ dR_t &= r R_t dt, \end{aligned}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, $r \geq 0$ are all known constants.

At each time $t \in [0, T]$, the investor spends some money from their portfolio for consumption and rebalances the remaining portfolio in the stock and the bond. Due to the time inconsistency, we shall restrict the strategies to feedback controls. We denote by $u(t, x) : [0, T] \times \mathbb{R}_+ \mapsto \mathbb{R}$ the fraction of portfolio value invested in the stock and $c(t, x) : [0, T] \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ the fraction of portfolio value consumed per unit of time. Then the controlled portfolio value process $\{X_t^{u,c}\}_{t \in [0, T]}$ follows the dynamics

$$dX_t^{u,c} = (r + (\mu - r)u(t, X_t^{u,c}) - c(t, X_t^{u,c})) X_t^{u,c} dt + \sigma X_t u(t, X_t^{u,c}) dW_t. \quad (4.1)$$

We say an investment-consumption strategy $(u, c) := \{(u(t, x), c(t, x))\}_{(t,x) \in [0, T] \times \mathbb{R}}$ is admissible if both the investment strategy and the consumption strategy are uniformly bounded, i.e.,

$$|u(t, x)| + c(t, x) \leq C, \quad \text{for all } (t, x) \in [0, T] \times \mathbb{R}.$$

The set of all admissible investment-consumption strategies is denoted by $\mathcal{A}_0$. Under any admissible investment-consumption strategy, the coefficients of the stochastic differential equation (SDE) (4.1) satisfy the linear growth and Lipschitz continuity conditions, thus the SDE admits a unique càdlàg solution that is square-integrable, i.e.,

$$\mathbb{E} \left[\sup_{t \in [0, T]} (X_t^{u,c})^2 \right] < \infty.$$

We denote the discounting function by $\beta_t : [0, T] \mapsto [0, 1]$, which is a non-increasing and differentiable function satisfying $\beta_0 = 1$. The objective of the optimal investment-consumption problem is to maximize the total discounted expected value of aggregated consumption and the terminal portfolio value under the log utility, that is,

$$\max_{(u,c) \in \mathcal{A}_0} J^{u,c}(t, x), \quad (4.2)$$

where

$$J^{u,c}(t, x) := \mathbb{E}_{t,x} \left[\int_t^T \beta_{s-t} \log(c(s, X_s^{u,c}) X_s^{u,c}) ds + \beta_{T-t} \log(X_T^{u,c}) \right]$$

is the objective function, and $\mathbb{E}_{t,x}[\cdot]$ denotes the conditional expectation $\mathbb{E}[\cdot | X_t = x]$.

In light of the non-exponential discounting, the optimal investment-consumption problem (4.2) has the issue of time inconsistency (Strotz, 1955). We adopt the time-inconsistent approach of (Björk et al., 2021) to define the equilibrium strategies. The intuition is that the time- t investor can only control her strategy within a small time period. Therefore, she aims to solve an optimal strategy that can be executed during this period, under the assumption that her future selves have already determined their own optimal strategies.

Definition 4.2.1 Let $(u^*, c^*) \in \mathcal{A}_0$ be an admissible investment-consumption strategy. For any $(t, x) \in [0, T] \times \mathbb{R}_+$, consider a perturbed strategy, $(u^\varepsilon, c^\varepsilon) := (u^\varepsilon(s, y), c^\varepsilon(s, y))_{(s, y) \in [t, T] \times \mathbb{R}_+}$, defined as

$$(u^\varepsilon(s, y), c^\varepsilon(s, y)) = \begin{cases} (\hat{u}, \hat{c}), & s \in [t, t + \varepsilon), \\ (u^*(s, y), c^*(s, y)) & s \in [t + \varepsilon, T], \end{cases}$$

where $\varepsilon > 0$ and $(\hat{u}, \hat{c}) \in \mathbb{R} \times \mathbb{R}_+$. Suppose that, for any $(t, x) \in [0, T] \times \mathbb{R}_+$ and any such perturbed strategy $(u^\varepsilon, c^\varepsilon)$, we have

$$\limsup_{\varepsilon \downarrow 0} \frac{J^{u^\varepsilon, c^\varepsilon}(t, x) - J^{u^*, c^*}(t, x)}{\varepsilon} \leq 0.$$

Then, (u^*, c^*) is an equilibrium investment-consumption strategy to problem (4.2), and $J^{u^*, c^*}(t, x)$ is the corresponding value function.

Through a formal verification argument, it can be shown that the value function of problem (4.2), i.e.,

$$V(t, x) = J^{u^*, c^*}(t, x), \quad (t, x) \in [0, T] \times \mathbb{R}_+,$$

and the auxiliary value function of problem (4.2), i.e.,

$$f(t, x, s) = \mathbb{E}_{t, x} \left[\int_t^T \beta_{l-s} \log(c^* X_l^{u^*, c^*}) dl + \beta_{T-s} \log X_T^{\pi^*} \right], \quad (t, x, s) \in [0, T] \times \mathbb{R}_+ \times [0, t],$$

solve the following system of extended Hamilton-Jacobi-Bellman (EHJB) equations

$$\begin{cases} \sup_{(u, c) \in \mathbb{R} \times \mathbb{R}_+} \{ \log(cx) + \mathcal{L}^{u, c} V(t, x) - \frac{\partial f}{\partial s}(t, x, t) \} = 0, & (t, x, s) \in [0, T] \times \mathbb{R}_+ \times [0, t], \\ \beta_{t-s} \log(c^* x) + \mathcal{L}^{u^*, c^*} f(t, x, s) = 0, & (t, x, s) \in [0, T] \times \mathbb{R}_+ \times [0, t], \\ V(T, x) = \log x, \quad f(T, x, s) = \beta_{T-s} \log x, & (x, s) \in \mathbb{R}_+ \times [0, T], \end{cases} \quad (4.3)$$

where (u^*, c^*) denotes the maximizer of the first equation, and the differential operator $\mathcal{L}^{u, c}$ is defined as

$$\mathcal{L}^{u, c} \varphi(t, x) := \frac{\partial \varphi}{\partial t}(t, x) + (r + (\mu - r)u - c)x \frac{\partial \varphi}{\partial x}(t, x) + \frac{1}{2} \sigma^2 u^2 x^2 \frac{\partial^2 \varphi}{\partial x^2}(t, x), \quad (4.4)$$

for any function $\varphi(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R}_+)$ that is first-order continuously differentiable in $t \in [0, T]$ and second-order continuously differentiable in $x \in \mathbb{R}_+$. We omit the proof of the verification theorem as it is a special case of Theorem 4.4.1 with exploration rate $\lambda = 0$.

By solving the system of EHJB equations (4.3), one obtains the equilibrium investment-consumption policy summarized in the following theorem. Note that Theorem 4.2.1 is a known result and can be found as Proposition 17.3 in (Björk et al., 2021).

Theorem 4.2.1 *An equilibrium investment-consumption policy to problem (4.2) is given by*

$$u^*(t, x) = \frac{\mu - r}{\sigma^2} \quad \text{and} \quad c^*(t, x) = \frac{1}{A(t)}, \quad (4.5)$$

where

$$A(t) = \beta_{T-t} + \int_t^T \beta_{T-l} dl. \quad (4.6)$$

The corresponding value function is given by

$$V(t, x) = J^{u^*, c^*}(t, x) = A(t) \log(x) + B_0(t),$$

where $B_0(t) = \int_t^T b_0(s) ds$ and

$$\begin{aligned} b_0(s) = & \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(s) - 1 + \log \left(\frac{1}{A(s)} \right) + \beta'_{T-s} \int_s^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{1}{A(l)} \right) dl \\ & + \int_s^T \beta'_{t-s} \int_s^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{1}{A(w)} \right) dw dl + \int_t^T \beta'_{t-s} \log \frac{1}{A(l)} dl. \end{aligned}$$

It is seen from Theorem 4.2.1 that under the log utility, the non-exponential discounting affects only the consumption, not the investment strategy. This result holds even for some stochastic discounting models; see (Harris & Laibson, 2013; Zou et al., 2014). However, we will see in the next section that the discounting factor will play a role in both the investment and the consumption policies in the exploratory formulation with RL.

At the end of this section, we study the impact of non-exponential discounting on consumption by comparing it with exponential discounting. To this end, we consider the class of *weighted discounting functions* proposed in (Ebert et al., 2020), defined as,

$$\beta_t = \int_0^\infty e^{-rt} dF(r),$$

where F is a cumulative distribution function supported on $[0, \infty)$. As elaborated in (Ebert et al., 2020), weighted discounting functions can model both intra-personal disagreement, when an investor is uncertain about which discount rate to use, and inter-personal disagreement, when a group of individuals have heterogeneous discount rates. It encompasses many widely-used non-exponential discounting models as special cases, including pseudo-exponential discounting (Karp, 2007; Ekeland & Pirvu, 2008), Gamma discounting (Loewenstein & Prelec, 1992; Weitzman, 2001), and uniform-uncertainty discounting (Sozou, 1998).

The following proposition compares the consumption strategies under an exponential discounting function and a weighted discounting function. The proof is provided in Section A.3.1 of the Appendix.

Proposition 4.2.1 *Consider an exponential discount function $\beta_t^{exp} := e^{-\rho t}$ with $\rho > 0$, and a non-exponential discounting function*

$$\beta_t^{non-exp} := \int_0^\infty e^{-rt} dF(r),$$

for some cumulative distribution function F supported on $[0, \infty)$, satisfying $\int_0^\infty r dF(r) = \rho$. We denote by $c^{*,exp}(t)$ and $c^{*,non-exp}(t)$ the corresponding equilibrium consumption strategies under β_t^{exp} and $\beta_t^{non-exp}$, respectively. Then, for any $t \in [0, T]$,

$$c^{*,exp}(t) \geq c^{*,non-exp}(t).$$

From the intra-personal perspective, the non-exponential discounting function $\beta_t^{non-exp}$ can be interpreted as the case where an investor is uncertain about which discount rate to apply. The condition

$$\int_0^\infty r dF(r) = \rho$$

ensures that the expected value of this “uncertain” discount rate coincides with the constant discount rate ρ in the exponential discounting, making the comparison fair from the perspective of expected value. The key message of Proposition 4.2.1 is that an investor facing *uncertainty* about the discount rate tends to *reduce consumption*.

4.3 Exploratory Investment-Consumption Problem

To the end of developing an RL algorithm for exploring situations in which the distribution of stock returns is uncertain to the investor, we now consider an exploratory formulation of the optimal investment–consumption problem with non-exponential discounting.

As in Section 4.2, let $\{W_t\}_{t \in [0, T]}$ be a one-dimensional Brownian Motion defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P} = \{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$ that satisfies the usual conditions. Denote by $\mathcal{P}(\mathbb{R} \times \mathbb{R}_+)$ the set of joint probability density functions over the space $\mathbb{R} \times \mathbb{R}_+$. Let $\pi : (t, x) \in [0, T] \times \mathbb{R}_+ \rightarrow \pi_{t,x} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)$ be a given stochastic feedback policy.

Correspondingly, the marginal density for the investment policy and the marginal density for the consumption policy are denoted by

$$\pi_{t,x}^U(u) = \int_{\mathbb{R}_+} \pi_{t,x}(u, c) dc \in \mathcal{P}(\mathbb{R}) \quad \text{and} \quad \pi_{t,x}^C(c) = \int_{\mathbb{R}} \pi_{t,x}(u, c) du \in \mathcal{P}(\mathbb{R}_+),$$

respectively. Applying the policy π to (4.1), we have that the corresponding controlled portfolio value process is given by

$$dX_t^\pi = [r + (\mu - r)\mathbf{u}_t^\pi - \mathbf{c}_t^\pi] X_t^{u,c} dt + \sigma X_t \mathbf{u}_t^\pi dW_t, \quad (4.7)$$

where the proportion of portfolio value in the stock $\mathbf{u}_t^\pi$ and the proportional consumption rate $\mathbf{c}_t^\pi$ are sampled from the joint density $\pi_{t,X_t^\pi}(u, c)$, i.e., the investment-consumption policy.

Since the investment and consumption actions are randomly sampled from the investment-consumption policy, we shall enlarge the filtered probability space. Following (Jia & Zhou, 2022b, 2023), we assume that the probability space is rich enough to support independent copies of a random vector that is uniformly distributed on $[0, 1]^2$. The new filtered probability space is denoted by $(\Omega, \mathcal{F}, \{\mathcal{G}_t\}_{t \in [0, T]}, \bar{\mathbb{P}})$, where $\mathcal{G}_t$ is the sigma-algebra generated by $\mathcal{F}_t$ and the copies of the uniform random vector up to time t , and $\bar{\mathbb{P}}$ is the product extension from $\mathbb{P}$. By the same law of large numbers argument as in (H. Wang & Zhou, 2020; H. Wang et al., 2020), if we randomly sample trajectories of $\mathbf{u}_t^\pi$ and $\mathbf{c}_t^\pi$ and take the average over the trajectories, the self-financing portfolio value process (4.7) follows the same distribution as

$$dX_t^\pi = \left(r + \int_{\mathbb{R}} (\mu - r) u \pi_t^U(u) du - \int_{\mathbb{R}_+} c \pi_t^C(c) dc \right) X_t^\pi dt + X_t^\pi \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi_t^U(u) du} dW_t. \quad (4.8)$$

For ease of notation, we write the marginal investment density as $\pi_t^U(u) := \pi_{t,X_t^\pi}^U(u)$ and the marginal consumption density as $\pi_t^C(c) := \pi_{t,X_t^\pi}^C(c)$.

With entropy regularization and a positive exploration weight $\lambda > 0$, the reward function of the exploratory investment-consumption problem under a given policy π is defined as

$$\begin{aligned} & J^\pi(t, x) \\ = & \mathbb{E}_{t,x}^{\bar{\mathbb{P}}} \left[\int_t^T \beta_{s-t} \left(\log(\mathbf{c}_s^\pi X_s^\pi) - \lambda \log(\pi_{s,X_s^\pi}(\mathbf{u}_s^\pi, \mathbf{c}_s^\pi)) \right) ds + \beta_{T-t} \log X_T^\pi \right] \end{aligned}$$

$$= \mathbb{E}_{t,x} \left[\int_t^T \beta_{s-t} \left(\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_s^\pi) - \lambda \log(\pi_s(u, c))) \pi_s(u, c) dc du \right) ds + \beta_{T-t} \log X_T^\pi \right]. \quad (4.9)$$

Recall that the entropy of a probability density function $\pi(x)$ is given by

$$\mathcal{H}(\pi) = - \int_{\mathbb{R}} \pi(x) \log(\pi(x)) dx.$$

The entropy term in the reward function (4.9) serves to encourage exploration within the control space. Therefore, the exploratory investment-consumption problem with non-exponential discounting is formulated as

$$\sup_{\pi \in \mathcal{A}} J^\pi(t, x). \quad (4.10)$$

We next define the set of admissible policies.

Definition 4.3.1 (Admissible Policy) *We say that $\pi := \{\pi_{t,x}(\cdot)\}_{t,x \in [0,T] \times \mathbb{R}_+}$ is an admissible policy if the following conditions are satisfied.*

- (1) *For any $(t, x) \in [0, T] \times \mathbb{R}_+$, $\pi_{t,x}(u, c) \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)$;*
- (2) *For the marginal investment policy, $\int_{\mathbb{R}} u^2 \pi_{t,x}^U(u) du < C_1$ for any $(t, x) \in [0, T] \times \mathbb{R}_+$;*
- (3) *For the marginal consumption policy, $\int_{\mathbb{R}_+} c \pi_{t,x}^C(c) dc < C_2$ for any $(t, x) \in [0, T] \times \mathbb{R}_+$;*
- (4) *$\int_{\mathbb{R}} \int_{\mathbb{R}_+} \pi_{t,x}(u, c) |\log \pi_{t,x}(u, c)| dc du < C_3(1 + |x|^2)$ for any $(t, x) \in [0, T] \times \mathbb{R}_+$.*

Here, C_1, C_2, C_3 are uniform constants that are independent of t and x . The set of all admissible policies is denoted by $\mathcal{A}$.

Under an admissible policy π , it is seen that the SDE for the controlled portfolio value process (4.8) satisfies the linear growth and Lipschitz continuity conditions, and thus admits a unique càdlàg solution that is square-integrable, i.e.,

$$\mathbb{E} \left[\sup_{t \in [0, T]} (X_t^\pi)^2 \right] < \infty.$$

Moreover, the conditions of admissible policies also imply that the reward function

$$J^\pi(t, x) < \infty,$$

for any $\pi \in \mathcal{A}$ and $(t, x) \in [0, T] \times \mathbb{R}_+$.

Inherited from the classical case without exploration, the issue of time inconsistency arises here due to non-exponential discounting. It is worth noting that time-inconsistent exploratory stochastic control problems have only recently begun to attract attention. Dai et al. (2023) studied an exploratory optimal investment problem under the mean-variance criterion, in which the time inconsistency stems from the variance term in the reward function. In contrast, our source of time inconsistency arises from the investor's time preferences. Despite the different origins of time inconsistency, we adopt a similar approach to that of (Dai et al., 2023) in defining equilibrium policies for Problem (4.10).

Definition 4.3.2 (Equilibrium Policy) *Let $\pi^* \in \mathcal{A}$ be an admissible investment-consumption policy. For any $(t, x) \in [0, T] \times \mathbb{R}_+$, consider a perturbed policy $\pi^\varepsilon := \{\pi_{s,y}^\varepsilon(\cdot)\}_{(s,y) \in [t,T] \times \mathbb{R}_+}$, defined as*

$$\pi_{s,y}^\varepsilon(u, c) = \begin{cases} v(u, c), & (s, y) \in [t, t + \varepsilon) \times \mathbb{R}_+, \\ \pi_{s,y}^*(u, c), & (s, y) \in [t + \varepsilon, T] \times \mathbb{R}_+, \end{cases}$$

where $\varepsilon > 0$, and $v(\cdot) \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)$ is an arbitrary density function satisfying

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (u^2 + c) v(u, c) dc du < \infty \quad \text{and} \quad \int_{\mathbb{R}} \int_{\mathbb{R}_+} v(u, c) |\log v(u, c)| dc du < \infty,$$

which implies $\pi^\varepsilon \in \mathcal{A}$. Suppose that for any $(t, x) \in [0, T] \times \mathbb{R}_+$ and any such perturbed strategy π^ε , we have

$$\limsup_{\varepsilon \downarrow 0} \frac{J^{\pi^\varepsilon}(t, x) - J^{\pi^*}(t, x)}{\varepsilon} \leq 0.$$

Then, π^* is an equilibrium investment-consumption policy to Problem (4.10), and $J^{\pi^*}(t, x)$ is the corresponding value function.

4.4 Theoretical Results

In this section, we present several theoretical results for the exploratory investment-consumption problem (4.10), including a verification theorem, an explicit characterization of the equilibrium investment-consumption policy, and a policy iteration theorem demonstrating that the equilibrium policy can be obtained through policy iteration starting from any time- and state-homogeneous admissible policy.

For any $\varphi(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R}_+)$ and $\pi \in \mathcal{A}$, we define the differential operator

$$\begin{aligned} \mathcal{L}^\pi \varphi(t, x) &:= \frac{\partial \varphi}{\partial t} + \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu - r)u - c) \pi_{t,x}(u, c) dc du \right) x \frac{\partial \varphi}{\partial x} \\ &\quad + \frac{x^2}{2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \sigma^2 u^2 \pi_{t,x}(u, c) dc du \frac{\partial^2 \varphi}{\partial x^2}. \end{aligned} \quad (4.11)$$

It can be seen that

$$\mathcal{L}^\pi \varphi(t, x) = \int_{\mathbb{R}} \int_{\mathbb{R}_+} \mathcal{L}^{u,c} \varphi(t, x) \pi_{t,x}(u, c) dc du,$$

where the differential operator $\mathcal{L}^{u,c}$ is defined in (4.4).

Next we present a formal verification theorem showing that the solution to a system of exploratory EHJB equations coincides with the value function of Problem (4.10). The proof is provided in Appendix A.3.2.

We say $f(t, x, s) \in \mathbb{C}^{1,2,1}([0, T] \times \mathbb{R}_+ \times [0, t])$ if the function is first-order continuously differentiable in t , second-order continuously differentiable in x , and first-order continuously differentiable in s .

Theorem 4.4.1 (Verification Theorem) *Suppose that there exist $V(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R}_+)$ and $f(t, x, s) \in \mathbb{C}^{1,2,1}([0, T] \times \mathbb{R}_+ \times [0, t])$ satisfying the following conditions.*

(1) *For any $(t, x) \in [0, T] \times \mathbb{R}$,*

$$\max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^\pi V(t, x) - \frac{\partial f}{\partial s}(t, x, t) \right\} = 0. \quad (4.12)$$

(2) *Let $\pi^* := \{\pi_{t,x}^*(u, c)\}_{(t,x) \in [0, T] \times \mathbb{R}_+}$ denote the maximizer attaining the supremum in (4.12) and suppose it satisfies, for any $(t, x, s) \in [0, T] \times \mathbb{R} \times [0, t]$,*

$$\beta_{t-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi_{t,x}^*(u, c))) \pi_{t,x}^*(u, c) dc du + \mathcal{L}^{\pi^*} f(t, x, s) = 0. \quad (4.13)$$

(3) *For any $x \in \mathbb{R}$,*

$$V(T, x) = \log x \quad \text{and} \quad f(T, x, s) = \beta_{T-s} \log x. \quad (4.14)$$

Then π^* is an equilibrium policy distribution to problem 4.10. Moreover, we have

$$V(t, x) = J^{\pi^*}(t, x) \quad (4.15)$$

and

$$f(t, x, s) = \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^{\pi^*} + \int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right]. \quad (4.16)$$

To solve for the equilibrium investment-consumption policy, we need the following lemma. Its proof is provided in Appendix A.3.3.

Lemma 4.4.0.1 *The unique maximizer for the problem*

$$\max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(c) - \lambda \log(\pi(u, c)) + \left((\mu - r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) \right) \pi(u, c) dc du, \quad (4.17)$$

where $A(t)$ is any deterministic positive function of $t \in [0, T]$, is given by

$$\pi^*(u, c) = \pi^{*,U}(u) \cdot \pi^{*,C}(c),$$

where

$$\pi^{*,U}(u) \sim \text{Normal} \left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A(t)} \right) \quad \text{and} \quad \pi^{*,C}(c) \sim \text{Gamma} \left(\frac{\lambda + 1}{\lambda}, \frac{A(t)}{\lambda} \right).^3 \quad (4.18)$$

The following theorem presents the equilibrium investment-consumption policy as well as the corresponding value function. Its proof is relegated to Appendix A.3.4.

Theorem 4.4.2 (Equilibrium Policy and Value Function) *An equilibrium investment-consumption policy for Problem (4.10) is given by $\pi_{t,x}^*(u, c) = \pi_{t,x}^{*,U}(u) \cdot \pi_{t,x}^{*,C}(c)$, where, for any $(t, x) \in [0, T] \times \mathbb{R}_+$,*

$$\pi_{t,x}^{*,U}(u) \sim \text{Normal} \left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A(t)} \right) \quad \text{and} \quad \pi_{t,x}^{*,C}(c) \sim \text{Gamma} \left(\frac{\lambda + 1}{\lambda}, \frac{A(t)}{\lambda} \right),$$

³The probability density function of $\text{Gamma}(\alpha, \beta)$ is given by $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

are respectively the marginal investment and consumption policies, and

$$A(t) = \beta_{T-t} + \int_t^T \beta_{T-l} dl. \quad (4.19)$$

The corresponding value function is given by

$$V(t, x) = A(t) \log(x) + B(t),$$

where $B(t) = \int_t^T b(s) ds$ with

$$\begin{aligned} b(s) = & \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(s) - (\lambda + 1) \log \left(\frac{A(s)}{\lambda} \right) + \lambda \log \left(\Gamma(1 + \frac{1}{\lambda}) \right) + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(s)} \right) \\ & + \beta'_{T-s} \int_s^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl + \int_s^T \beta'_{l-s} \int_s^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\ & + \int_s^T \beta'_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma(1 + \frac{1}{\lambda}) \right) \right) dl. \end{aligned} \quad (4.20)$$

Moreover, the auxiliary value function is given by

$$\begin{aligned} f(t, x, s) = & \left(\beta_{T-s} + \int_t^T \beta_{l-s} dl \right) \log(x) + \beta_{T-s} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl \\ & + \int_t^T \beta_{l-s} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\ & + \int_t^T \beta_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma(1 + \frac{1}{\lambda}) \right) \right) dl. \end{aligned} \quad (4.21)$$

We see from Theorem 4.4.2 that the equilibrium policy π^* is a product of the marginal investment policy $\pi_{t,x}^{*,U}$ and the marginal consumption policy $\pi_{t,x}^{*,C}$, indicating independence between investment and consumption. Mathematically, this follows from the separability of the two control variables u and c in the EHJB (4.12); see also equation (A.38) in the Appendix. Note that both controls are proportional to the wealth process. Therefore, the independence refers only to the ratios (of investment and consumption) relative to the realized wealth at each point in time. Moreover, it is well known that the HJB approach,

together with standard verification arguments, does not establish uniqueness of the optimal or equilibrium strategies. Therefore, this result does not exclude the possibility of a dependent policy that is nevertheless equivalently optimal.

The following corollary shows that, if exploration is not encouraged, that is, when the exploration weight $\lambda \rightarrow 0$, the equilibrium policies in Theorem 4.4.2 reduce to the equilibrium strategies of the classical problem (4.2) without exploration. Its proof is relegated to Appendix A.3.5.

Corollary 4.4.1 *As the exploration weight $\lambda \rightarrow 0$, the equilibrium policies of the exploratory investment-consumption problem in Theorem 4.4.2 converge to the equilibrium strategy of the classical non-exploratory problem in Theorem 4.2.1. In addition, the corresponding value functions also converge.*

Corollary 4.4.2 *Suppose we instead use exponential discounting, the equilibrium policy for investment and consumption will take the same form as in Theorem 4.4.2, with*

$$A(t) = \frac{1}{\gamma}(1 - e^{-\gamma(T-t)}) + e^{-\gamma(T-t)}. \quad (4.22)$$

The following proposition provides a comparison similar to Proposition 4.2.1, namely, the equilibrium investment-consumption policies under an exponential discounting function and under a weighted discounting function with the same mean. This setting can be viewed as the case of an investor facing a certain discount factor versus an uncertain discount factor of the same mean. Its proof follows immediately from Theorem 4.4.2 and Proposition 4.2.1.

Proposition 4.4.1 *Consider an exponential discounting function $\beta_t^{exp} := e^{-\rho t}$ with $\rho > 0$, and a non-exponential discounting function*

$$\beta_t^{non-exp} := \int_0^\infty e^{-rt} dF(r),$$

for some cumulative distribution function F supported on $[0, \infty)$, satisfying $\int_0^\infty r dF(r) = \rho$. The corresponding investment-consumption policies are given by

$$\begin{aligned} \pi_{t,x}^{*,U,exp}(u) &\sim \text{Normal} \left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A^{exp}(t)} \right) \\ \pi_{t,x}^{*,C,exp}(c) &\sim \text{Gamma} \left(\frac{\lambda + 1}{\lambda}, \frac{A^{exp}(t)}{\lambda} \right), \end{aligned}$$

and,

$$\begin{aligned}\pi_{t,x}^{*,U,non-exp}(u) &\sim \text{Normal}\left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A^{non-exp}(t)}\right) \\ \pi_{t,x}^{*,C,non-exp}(c) &\sim \text{Gamma}\left(\frac{\lambda + 1}{\lambda}, \frac{A^{non-exp}(t)}{\lambda}\right),\end{aligned}$$

such that, for any $t \in [0, T]$,

$$A^{exp}(t) := \beta_{T-t}^{exp} + \int_t^T \beta_{T-l}^{exp} dl \leq \beta_{T-t}^{non-exp} + \int_t^T \beta_{T-l}^{non-exp} dl := A^{non-exp}(t).$$

A noticeable distinction between the exploratory formulation with RL (Theorem 4.4.2) and the classical formulation without RL (Theorem 4.2.1) is that the discounting function now affects not only the consumption policy but also the investment policy. Moreover, Proposition 4.4.1 implies that when an investor faces *uncertainty* about the discount rate, she tends to adopt a more *conservative* investment-consumption policy in the sense that the Gaussian-distributed investment policy has the same mean but a lower variance, and the Gamma-distributed consumption policy has both a lower mean and lower variance.⁴

In the end of this section, we provide a policy iteration theorem. Standard policy iteration in RL consists of two steps: For $i = 0, 1, 2, \dots$,

1. Policy evaluation: Start with a policy $\pi^{(i)}$ and calculate its corresponding value function $V^{(i)}$.
2. Policy improvement: Update the policy $\pi^{(i+1)}$ according to a prescribed iteration rule and show that $V^{(i+1)} \geq V^{(i)}$.

However, as noted by (Dai et al., 2023), such monotone improvement, i.e., $V^{(i+1)} \geq V^{(i)}$ may not hold for problems with time inconsistency. For our exploratory investment-consumption problem with non-exponential discounting, in light of the EHJB equation (4.12), we propose the following policy iteration procedure consisting of two steps: For $i = 0, 1, 2, \dots$,

⁴The mean and variance of a $\text{Gamma}(\alpha, \beta)$ distribution are given by $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively.

1. Policy evaluation: Start with a policy $\pi^{(i)}$ and calculate its corresponding value functions

$$\begin{aligned}
V^{(i)}(t, x) &:= J^{\pi^{(i)}}(t, x) \\
&= \mathbb{E}_{t,x} \left[\int_t^T \beta_{s-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_s^{\pi^{(i)}}) - \lambda \log(\pi_s^{(i)}(u, c)) \right) \pi_s^{(i)}(u, c) dc du ds \right. \\
&\quad \left. + \beta_{T-t} \log X_T^{\pi^{(i)}} \right],
\end{aligned}$$

and

$$\begin{aligned}
f^{(i)}(t, x, s) &= \mathbb{E}_{t,x} \left[\int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_l^{\pi^{(i)}}) - \lambda \log(\pi_l^{(i)}(u, c)) \right) \pi_l^{(i)}(u, c) dc du dl \right. \\
&\quad \left. + \beta_{T-s} \log X_T^{\pi^{(i)}} \right].
\end{aligned}$$

2. Policy iteration: Update the policy to $\pi^{(i+1)}$ according to the EHJB equation (4.12),

$$\begin{aligned}
\pi^{(i+1)} &= \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^\pi V^{(i)}(t, x) \right. \\
&\quad \left. - \frac{\partial f^{(i)}}{\partial s}(t, x, t) \right\} \\
&= \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(c) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^\pi V^{(i)}(t, x) \right\}.
\end{aligned} \tag{4.23}$$

The following theorem shows that, starting from an arbitrary time- and state-homogeneous admissible policy $\pi^{(0)}$, the above policy iteration procedure (4.23) can converge to the equilibrium policy in Theorem 4.4.2 in one step. This finding is consistent with Theorem 4.2 of (Dai et al., 2023) in the absence of market state variable. The proof is provided in Appendix A.3.6.

Theorem 4.4.3 (Policy Iteration Theorem) *Consider an arbitrary time- and state-homogeneous admissible policy $\pi_{t,x}^{(0)} \equiv \pi^{(0)} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)$ for any $(t, x) \in [0, T] \times \mathbb{R}_+$. Let*

$V^{(0)}(t, x) = J^{\pi^{(0)}}(t, x)$ be the corresponding value function and

$$f^{(0)}(t, x, s) = \mathbb{E}_{t,x} \left[\int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_l^{\pi^{(0)}}) - \lambda \log(\pi_l^{(0)}(u, c)) \right) \pi_l^{(0)}(u, c) dc du dl \right. \\ \left. + \beta_{T-s} \log X_T^{\pi^{(0)}} \right]$$

be the corresponding auxiliary value function. We iterate the policy by

$$\pi^{(1)} = \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(c) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^{\pi} V^{(0)}(t, x) \right\}.$$

Then we have

$$\pi^{(1)} = \pi^*,$$

where π^* is the equilibrium investment-consumption policy in Theorem 4.4.2.

4.5 RL Algorithm

In this section, we design RL algorithms to iteratively learn the equilibrium policies for investment and consumption without requiring statistical estimation or prior knowledge of the true market parameters. Specifically, we design two RL algorithms: one based on the policy evaluation approach, which proposes a loss function following from the martingale property of the analytical solution in Theorem 4.4.2, and the other one based on the q-learning approach (Jia & Zhou, 2023). While the loss functions from both approaches involve the orthogonality conditions of different martingales, we refer to the loss function derived from our analytical solution as the *Orthogonality Condition (OC) loss*, and we refer to the loss function derived from (Jia & Zhou, 2023) as the q-learning loss.

For clarity, we repeat the stock and bond dynamics

$$dS_t = \mu_{true} S_t dt + \sigma_{true} S_t dW_t \\ dR_t = r R_t dt,$$

where the true growth rate and volatility of the stock, μ_{true} and σ_{true} , are unknown, and the risk-free interest rate r is known.

We parameterize the investment and consumption policies with the market parameters $\theta := (\mu, \sigma)$. It follows from (4.18) that the parameterized $\pi_{t,x}^\theta(u, c) = \pi_{t,x}^{\theta,U}(u) \cdot \pi_{t,x}^{\theta,C}(c)$ with

$$\pi_{t,x}^{\theta,U}(u) \sim \text{Normal}\left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A(t)}\right) \quad \text{and} \quad \pi_{t,x}^{\theta,C}(c) \sim \text{Gamma}\left(\frac{\lambda + 1}{\lambda}, \frac{A(t)}{\lambda}\right) \quad (4.24)$$

where $A(t) = \beta_{T-t} + \int_t^T \beta_{T-s} ds$ is identical to the expression in (4.19). Note that $A(t)$ does not depend on the market parameter θ , and it is explicitly determined when the discounting function is specified. Under the parameterized policy π^θ , the corresponding wealth process $\{X_t^{\pi^\theta}\}_{t \in [0, T]}$ follows the dynamics

$$\begin{aligned} dX_t^{\pi^\theta} &= \left[r + (\mu_{true} - r)u_t^{\pi^\theta} - c_t^{\pi^\theta} \right] X_t^{\pi^\theta} dt + X_t^{\pi^\theta} \sigma_{true} u_t^{\pi^\theta} dW_t \\ &= \left(r + \int_{\mathbb{R}} [(\mu_{true} - r)u - c] \pi_t^{\theta,U}(u) du - \int_{\mathbb{R}_+} c \pi_t^{\theta,C}(c) dc \right) X_t^{\pi^\theta} dt \\ &\quad + X_t^{\pi^\theta} \sqrt{\int_{\mathbb{R}} \sigma_{true}^2 u^2 \pi_t^{\theta,U}(u) du} dW_t \\ &= \left(r + \frac{(\mu_{true} - r)(\mu - r)}{\sigma^2} - \frac{\lambda + 1}{A(t)} \right) X_t^{\pi^\theta} dt + X_t^{\pi^\theta} \sigma_{true} \sqrt{\frac{(\mu - r)^2}{\sigma^4} + \frac{\lambda}{\sigma^2 A(t)}} dW_t. \end{aligned} \quad (4.25)$$

The corresponding value function can be parameterized as

$$V^\theta(t, x) = A(t) \log(x) + B^\theta(t), \quad (4.26)$$

where $B^\theta(t) = \int_t^T b^\theta(s) ds$ with $b^\theta(s)$ given by (4.20), i.e.,

$$\begin{aligned} b^\theta(s) &= \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(s) - (\lambda + 1) \log\left(\frac{A(s)}{\lambda}\right) + \lambda \log\left(\Gamma\left(1 + \frac{1}{\lambda}\right)\right) \\ &\quad + \frac{\lambda}{2} \log\left(\frac{2\pi\lambda}{\sigma^2 A(s)}\right) + \beta'_{T-s} \int_s^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl \\ &\quad + \int_s^T \beta'_{l-s} \int_s^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\ &\quad + \int_s^T \beta'_{l-s} \left(\frac{\lambda}{2} \log\left(\frac{2\pi\lambda}{\sigma^2 A(l)}\right) - (\lambda + 1) \log\left(\frac{A(l)}{\lambda}\right) + \frac{3\lambda + 2}{2} + \lambda \log\left(\Gamma\left(1 + \frac{1}{\lambda}\right)\right) \right) dl. \end{aligned} \quad (4.27)$$

Moreover, by (4.21), we denote the auxiliary value function following the parameterized policy π^θ by

$$\begin{aligned}
f^\theta(t, x, s) = & \left(\beta_{T-s} + \int_t^T \beta_{l-s} dl \right) \log(x) + \beta_{T-s} \int_t^T \left(r + \frac{(\mu-r)^2}{2\sigma^2} - \frac{3\lambda+2}{2A(l)} \right) dl \\
& + \int_t^T \beta_{l-s} \int_t^l \left(r + \frac{(\mu-r)^2}{2\sigma^2} - \frac{3\lambda+2}{2A(w)} \right) dw dl \\
& + \int_t^T \beta_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda+1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda+2}{2} + \lambda \log(\Gamma(\frac{\lambda+1}{\lambda})) \right) dl.
\end{aligned} \tag{4.28}$$

4.5.1 Policy Evaluation Approach

In light of the EHJB (4.12), we introduce the process, for $t \in [0, T]$,

$$\begin{aligned}
M_t^\theta := & V^\theta(t, X_t^{\pi^\theta}) - \int_0^t \left(- \int_{\mathbb{R}_+} \log(cX_s^{\pi^\theta}) \pi_s^{\theta,C}(c) dc \right. \\
& \left. + \lambda \int_{\mathbb{R}} \int_{\mathbb{R}_+} \pi_s^\theta(u, c) \log \pi_s^\theta(u, c) dc du + \frac{\partial f^\theta}{\partial s}(s, X_s^{\pi^\theta}, s) \right) ds.
\end{aligned} \tag{4.29}$$

The proof of the following theorem is postponed to Appendix A.3.7.

Theorem 4.5.1 *The process $\{M_t^\theta\}_{t \in [0, T]}$ is an $(\{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$ martingale when $\theta = (\mu, \sigma) := (\mu_{true}, \sigma_{true})$.*

Moreover, we know that any square-integrable martingale $M := \{M_t\}_{t \in [0, T]}$ has an orthogonality condition

$$\mathbb{E} \left[\int_0^T \zeta_t dM_t \right] = 0,$$

where the test function $\zeta := \{\zeta_t\}_{t \in [0, T]}$ is an arbitrary $\{\mathcal{F}_t\}_{t \in [0, T]}$ -adapted process that is square-integrable with respect to M . This, together with the martingale property of M^θ , allow us to define an Orthogonality Condition Loss function. Inspired by (Jia & Zhou, 2022a), we choose the test function as the partial derivative of the parameterized value function V^θ with respect to its parameters θ .

Definition 4.5.1 (Orthogonality Condition (OC) Loss) *Choose*

$$\eta_t := \frac{\partial V^\theta(t, x)}{\partial \theta} = \left(\frac{\partial V^\theta(t, x)}{\partial \theta_1}, \frac{\partial V^\theta(t, x)}{\partial \theta_2} \right),$$

where $(\theta_1, \theta_2) := (\mu, \sigma)$. For $(t, x) \in [0, T] \times \mathbb{R}$ and $j = 1, 2$, the OC loss is given by

$$L_{OC}(\theta_j) = \mathbb{E}_{t,x} \left[\int_t^T \frac{\partial V^\theta(s, X_s^{\pi^\theta})}{\partial \theta_j} dM_s^\theta \right].$$

The discretization is given by

$$L_{OC}(\theta_j) = \sum_{n=0}^{N-1} \frac{\partial V^\theta(t_n, X_{t_n}^{\pi^\theta})}{\partial \theta_j} (M_{t_{n+1}}^\theta - M_{t_n}^\theta), \quad (4.30)$$

where, by (4.26) and (4.27),

$$\begin{aligned} \frac{\partial V^\theta(t, x)}{\partial \theta_1} &= \frac{\partial V^\theta(t, x)}{\partial \mu} = \frac{\mu - r}{\sigma^2} \int_t^T \left(A(s) + \beta'_{T-s}(T - s) + \int_s^T \beta'_{l-s}(l - s) dl \right) ds \quad (4.31) \\ \frac{\partial V^\theta(t, x)}{\partial \theta_2} &= \frac{\partial V^\theta(t, x)}{\partial \sigma} \\ &= \int_t^T \left(-\frac{(\mu - r)^2}{\sigma^3} \left(A(s) + \beta'_{T-s}(T - s) + \int_s^T \beta'_{l-s}(l - s) dl \right) - \frac{\lambda}{\sigma} \beta_{T-s} \right) ds. \end{aligned} \quad (4.32)$$

4.5.2 q-Learning Approach

The q-learning algorithm, introduced by (Jia & Zhou, 2023), is a continuous-time extension to the classical Q-Learning method (Sutton & Barto, 2018). While the Q-function quantifies the value of taking an action at a state following a policy, the q-function is the rate of improvement in the value function for taking an action at a state following a policy.

Consider an admissible policy $\pi \in \mathcal{A}$ and a perturbed policy that takes constant actions $(u, c) \in \mathbb{R} \times \mathbb{R}_+$ over $[t, t + \Delta t)$ and coincides with π on $[t + \Delta t, T]$ for some small time interval Δt . Then, the Q-function, $Q_{\Delta t}(t, x, (u, c); \pi)$, is the expected reward derived from the perturbed policy, i.e.,

$$Q_{\Delta t}(t, x, (u, c); \pi) := J^{\pi^{\Delta t}}(t, x),$$

where the policy $\pi^{\Delta t}$ is given by

$$\pi_{l,y}^{\Delta t}(\cdot) = \begin{cases} (u, c), & l \in [t, t + \Delta t), \\ \pi_{l,y}(\cdot), & l \in [t + \Delta t, T], \end{cases}$$

Note that $\pi^{\Delta t}$ can be viewed as a perturbed policy as in Definition 4.3.2, and over the short time interval $[t, t + \Delta t)$, the policy takes a deterministic investment-consumption control $(u, c) \in \mathbb{R} \times \mathbb{R}_+$.

Following the same derivation as for (A.36), we obtain

$$Q_{\Delta t}(t, x, (u, c); \pi) = J^\pi(t, x) + \Delta t \left(\log(cx) - \frac{\partial f^\pi}{\partial s}(t, x, t) + \mathcal{L}^{u,c} J^\pi(t, x) \right) + o(\Delta t), \quad (4.33)$$

where

$$f^\pi(t, x, s) = \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^\pi + \int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^\pi) - \lambda \log(\pi_l(u, c))) \pi_l(u, c) dc du dl \right]. \quad (4.34)$$

Note that the relation (4.33) parallels Proposition 3 in (Jia & Zhou, 2023). We recall from (4.4) that

$$\mathcal{L}^{u,c} J^\pi(t, x) = \frac{\partial J^\pi}{\partial t}(t, x) + (r + ((\mu - r)u - c)) x \frac{\partial J^\pi}{\partial x}(t, x) + \frac{1}{2} \sigma^2 u^2 x^2 \frac{\partial^2 J^\pi}{\partial x^2}(t, x).$$

Then, the q-function is denoted by

$$\begin{aligned} q(t, x, (u, c); \pi) &:= \lim_{\Delta t \rightarrow 0} \frac{Q_{\Delta t}(t, x, (u, c); \pi) - J^\pi(t, x)}{\Delta t} \\ &= \log(cx) - \frac{\partial f^\pi}{\partial s}(t, x, t) + \mathcal{L}^{u,c} J^\pi(t, x). \end{aligned} \quad (4.35)$$

By a similar argument as for Theorem 4.4.1, (4.12), and (4.35), we have that

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (q(t, x, (u, c); \pi) - \lambda \log(\pi_{t,x}(u, c))) \pi_{t,x}(u, c) dc du = 0,$$

which parallels Theorem 6 in (Jia & Zhou, 2023).

Let π^θ be a policy defined in (4.24) parameterized by $\theta = (\mu, \sigma)$. Let V^θ and f^θ be the corresponding value function and the auxiliary value function given in (4.26) and (4.28), respectively. Define a process $\{N_t^\theta\}_{t \in [0, T]}$ as

$$N_t^\theta := V^\theta(t, X_t^{\pi^\theta}) + \int_0^t \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_s^{\pi^\theta}) - q(s, X_s^{\pi^\theta}, (u, c); \pi^\theta) \right) \pi_s^\theta(u, c) dc du ds$$

$$-\int_0^t \frac{\partial f^\theta}{\partial s}(s, X_s^{\pi^\theta}, s) ds. \quad (4.36)$$

The proof of the following theorem is given in the Appendix [A.3.8](#).

Theorem 4.5.2 *The process $\{N_t^\theta\}_{t \in [0, T]}$ is an $(\{\mathcal{F}_t\}_{t \in [0, T]}, \mathbb{P})$ martingale when $\theta = (\mu, \sigma) := (\mu_{true}, \sigma_{true})$.*

Following the aforementioned parameterization, this theorem allows us to define a loss function by applying the orthogonality condition of the martingale N .

Definition 4.5.2 (q-Learning Loss) *Choose $\eta_t = \frac{\partial V^\theta(t, x)}{\partial \theta} = (\frac{\partial V^\theta(t, x)}{\partial \theta_1}, \frac{\partial V^\theta(t, x)}{\partial \theta_2})$, where $(\theta_1, \theta_2) = (\mu, \sigma)$. For $(t, x) \in [0, T] \times \mathbb{R}$ and $j = 1, 2$, the q-learning loss is given by*

$$L_{ql}(\theta_j) = \mathbb{E} \left[\int_t^T \frac{\partial V^\theta(l, X_l^\theta)}{\partial \theta_j} dN_l^\theta \right]. \quad (4.37)$$

With the same time discretization as in the last subsection, the q-learning loss can be approximated by

$$L_{ql}(\theta_j) = \sum_{n=0}^{N-1} \frac{\partial V^\theta(t_n, X_{t_n}^\theta)}{\partial \theta_j} (N_{t_{n+1}}^\theta - N_{t_n}^\theta), \quad (4.38)$$

for $j = 1, 2$ and the partial derivatives $\frac{\partial V^\theta(t_n, X_{t_n}^\theta)}{\partial \theta_j}$ are given in [\(4.31\)](#) and [\(4.32\)](#).

We summarize the whole training procedure in Algorithm [3](#).

Algorithm 3: Training with Orthogonality Condition Loss and q-Learning Loss

Initialize number of epochs N_{epochs} , the investment horizon T , the mesh size of the time step Δt , the exploration parameter $\lambda > 0$, the initial portfolio value $x_0 > 0$, the learning rates $\eta = (\eta_1, \eta_2)$, the “grounding true” market parameters

$\theta_{true} = (\theta_{true,1}, \theta_{true,2}) \equiv (\mu_{true}, \sigma_{true})$, model parameters

$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}) = (\mu^{(0)}, \sigma^{(0)})$.

Given the non-exponential discounting function $\beta(t)$.

Compute $N = \frac{T}{\Delta t}$.

for $k = 0, \dots, N_{epoch} - 1$ **do**

 Fix a realized path of the Brownian motion $\{W_{t_n}^{(k)}\}_{n=0}^N$, with

$\Delta W_{t_n}^{(k)} := W_{t_{n+1}}^{(k)} - W_{t_n}^{(k)} \sim N(0, \Delta t)$.

 Simulate $\{X_{t_n}^{(k)}\}_{n=0}^N$ via (4.25), with (μ, σ) replaced by $(\mu^{(k)}, \sigma^{(k)})$ and $\{W_{t_n}\}_{n=0}^N$ replaced by $\{W_{t_n}^{(k)}\}_{n=0}^N$.

 Compute $\{\frac{\partial V^{\theta^{(k)}}}{\partial \mu}(t_n, X_{t_n}^{(k)})\}_{n=0}^N$ and $\{\frac{\partial V^{\theta^{(k)}}}{\partial \sigma}(t_n, X_{t_n}^{(k)})\}_{n=0}^N$ via (4.31) and (4.32), with (μ, σ) replaced by $(\mu^{(k)}, \sigma^{(k)})$.

for $j = 1, 2$ **do**

 If training with orthogonal condition loss, compute

$L(\theta_j^{(k)}) := L_{oc}(\theta_j^{(k)}; \{X_{t_n}^{(k)}\}_{n=0}^N)$ via Eq (4.30).

 If training with q-learning loss, compute $L(\theta_j^{(k)}) := L_{ql}(\theta_j^{(k)}; \{X_{t_n}^{(k)}\}_{n=0}^N)$ via Eq (4.38).

 Update $\theta_j^{(k+1)} \leftarrow \theta_j^{(k)} + \eta_j \times L(\theta_j^{(k)})$

end

end

4.6 Numerical Results

In this section, we present the results of an investment-consumption problem based on simulation. Under two discounting settings, non-exponential discounting and exponential discounting, we train RL models independently using the two aforementioned loss functions, OC loss and q-learning loss, totaling of four setups. We demonstrate the model convergence of each setup and compare them under each discounting setting and loss function setting.

For simplicity of demonstration, we employ a gamma discounting function as the non-

exponential discounting function, i.e.,

$$\beta_t = (1 + \tilde{\alpha}t)^{-\frac{\tilde{\beta}}{\tilde{\alpha}}}, \quad (4.39)$$

where $\tilde{\alpha}, \tilde{\beta} > 0$ are constant parameters. Gamma discounting was first introduced by (Weitzman, 2001) as a way to address the challenge of uncertain future discount rates. For comparison purposes, we also consider the classical exponential discounting, i.e.,

$$\beta_t = e^{-\rho t},$$

where $\rho > 0$ is a constant parameter.

We consider an investor who decides the consumption budget and rebalances the portfolio on a monthly basis, over a one-year time horizon. Therefore, $T = 1$ and $\Delta t = \frac{1}{12}$. We generate synthetic stock prices from a market simulator, assuming the grounding true growth rate of the stock, μ_{true} , is 0.2 and the stock return volatility, σ_{true} , is 0.4, both annualized. The risk-free interest rate is fixed at 1% and is considered as given, as in practice people can use short-term government bond yield rate (such as the 3-month US treasury bill rate) as an approximation of the risk-free rate. We initialize the RL model with randomly picked $(\mu^{(0)}, \sigma^{(0)}) = (0.1, 0.5)$, and hide $(\mu_{true}, \sigma_{true})$ from the model. For the hyperparameters of the discounting functions, we adopt $\tilde{\alpha} = 0.0225, \tilde{\beta} = 0.04$ selected by (Weitzman, 2001) in our simulation study, which was constructed from the survey results based on the opinions of 2,160 economists. We also select $\rho = 0.0175$, which is the effective constant discounting rate of our selected gamma discounting function (Weitzman, 2001). The simulation study is done on a MacBook with Intel Core i7 2.6 GHz 6-core CPU and 16GB RAM, by CPU computation only.

Figure 4.1 shows the model training results under non-exponential discounting. Clearly, all parameters (μ and σ) converge to the corresponding true levels when training with OC loss and q-learning loss. This indicates that the RL model successfully calibrates the underlying dynamics of this synthetic market, through iteratively interacting (investing and consuming) within the market. The RL model converges in fewer epochs (around 700) when training with q-learning loss, compared to when training with OC loss (around 780 epochs). Yet, training with q-learning loss takes slightly longer time (around 220 seconds) than training with OC loss (around 210 seconds).

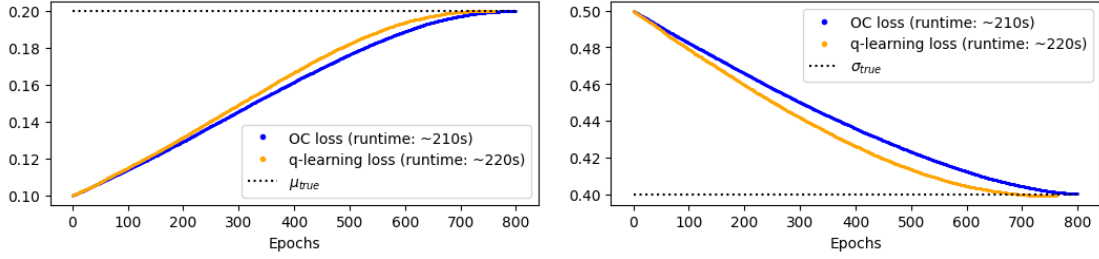


Figure 4.1: Parameter convergence using non-exponential discounting, μ (left) and σ (right).

Figure 4.2 illustrates the model training result under exponential discounting, where we find similar observations as in Figure 4.1. All parameters converge to their corresponding true values. Training with q-learning again takes fewer epochs (around 560) but slightly longer runtime to convergence (around 150 seconds), compared to that of training with OC loss (around 670 epochs in around 140 seconds). These consistent observations also demonstrate the robustness of our model against various types of risk aversions, as reflected by the discounting functions. Furthermore, the model converges faster under exponential discounting than under non-exponential discounting. This is expected because with exponential discounting, the time inconsistency no longer exists and the $\frac{\partial f}{\partial s}(t, x, t)$ term will disappear from the HJB equation (4.12) in Theorem 4.4.1, being replaced with $\rho V(t, x)$. As a result, the definitions of the M and N processes in Sections 4.5.1 and 4.5.2 will be reduced to simpler forms, making the computation easier at each epoch.

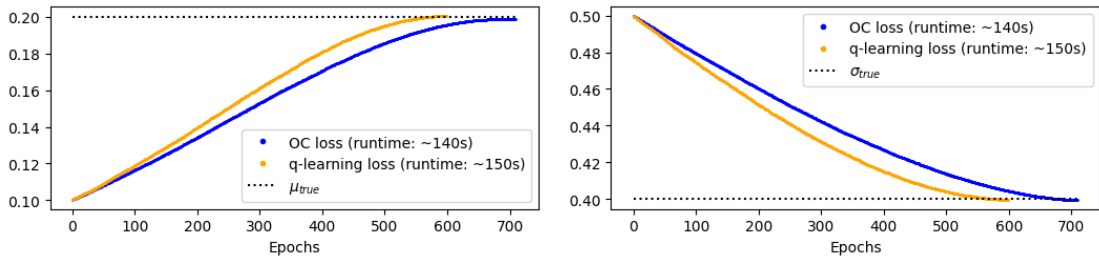


Figure 4.2: Parameter convergence using exponential discounting, μ (left) and σ (right).

4.7 Conclusion

In this chapter, we apply the equilibrium approach to solve an exploratory formulation of the optimal investment-consumption problem, with log utility and non-exponential discounting. We show that the investment and consumption are separable, hence our equilibrium investment and consumption policies are two independent probability distributions. Specifically, we solve for a Gaussian investment policy and a Gamma consumption policy. The domains of the equilibrium policies match the feasible ranges of values that investment and consumption can take. Investment, as defined by the proportion of the portfolio value invested in the risky asset, can be positive or negative, depending on the long and short position in the asset. Consumption, on the other hand, cannot be negative and the support of a Gamma distribution is $\mathbb{R}_+$.

Additionally, we note that our choice of discounting function influences our equilibrium investment and consumption policies, as the discounting function appears in the denominators of the mean and variance of the Gamma consumption policy and the variance of the Gaussian investment policy. We show that the non-exponential discounting function is everywhere greater than the exponential discounting function, when the means of the discount rates are equal. Therefore, with non-exponential discounting, the mean and variance of the Gamma consumption policy, as well as the variance of the Gaussian investment policy will all decrease, despite the mean of the Gaussian investment policy remaining unaffected. We remark that this is because a non-exponential discounting investor faces future ambiguity in the discount rates, hence behaving more conservatively than an exponential discounting investor with deterministic discount rates. Moreover, our results under the exploratory setting generalize the classical non-exploratory setting. Under the classical setting, the discounting function only appears in the optimal consumption policy. Hence an investor with non-exponential discounting will only reduce consumption, leaving the investment unchanged. Under our exploratory setting, however, an investor with non-exponential discounting will not only reduce the mean of the consumption policy but also reduce the variance of both the consumption and investment policies. As a result, they will cut consumption and restrain the exploration of both consumption and investment. Finally, we construct two martingales using our analytical solutions and leverage their properties to derive two loss functions — an orthogonality condition loss and a q-learning loss. Our RL training and parameter updating procedures do not involve estimation of parameters, hence avoiding the risk of estimation errors; see (Luenberger, 2013) for the *mean-blur problem*, a major argument against maximum likelihood estimation for mean returns of assets in finance. Our RL model learns the true parameter values directly from epochs of investment and consumption histories and iterative interactions with the market dynamics. Our

simulation study empirically justifies our RL algorithm and our choices of loss functions. In four independent training processes, using exponential and non-exponential discounting and the two loss functions, we observe clear and accurate parameter convergence from the RL models. Both loss functions converge in similar time for a given discounting function, whereas exponential discounting converges faster than non-exponential discounting for a given loss function.

Chapter 5

Equilibrium Exploratory Stopping Problem with Non-Exponential Discounting

5.1 Introduction

Optimal stopping problems aim to select the most rewarding time to terminate a game or exit from a dynamically evolving process, immediately realizing the termination payoff, when the cost of waiting surpasses the expected value of a potentially better future payoff. These problems appear throughout a broad range of disciplines, such as the secretary problem¹ studied in (Lindley, 1961; Chow et al., 1964; Ferguson, 1989; Hsiao & Yang, 2000; Bruss, 2000) and the optimal exercise of American put options² studied in (McKean Jr., 1965; Merton, 1973; Peskir & Shiryaev, 2006; Hull & Basu, 2016).

Conceptually, optimal stopping can be viewed as a specialized control problem where the agent has only two available actions: stop or continue. Built on the martingale approaches and dynamic programming (DP), the optimal stopping problem can be reduced to a *variational inequality*, or a *free boundary problem*. This involves characterization of

¹A recruiter chooses the best candidate by interviewing a randomly permuted group of candidates, giving the accept/reject decision one-by-one without the possibility to re-accept an already rejected candidate.

²An option holder decides when to exercise the right to buy or sell an asset at a prespecified price, on or before the maturity of the option.

the value function via a partial differential equation (PDE), while simultaneously determining the stopping and continuation regions (Friedman, 1975; Peskir & Shiryaev, 2006). When the model (i.e., the underlying dynamics of the stochastic processes) is fully known and supposing regularity conditions are satisfied, we can solve the free-boundary PDE numerically (if it is not analytically tractable) via finite difference methods; see for example (Forsyth & Vetzal, 2002; Achdou & Pironneau, 2005; Kloeden & Platen, 2018). However, such approaches suffer from a critical limitation — the assumption of full knowledge of the model, which is unrealistic when we are modeling a real-world dynamic process such as the price movement of a stock.

To handle this limitation, recent stochastic control literature has applied Reinforcement Learning (RL) methods in conjunction with dynamic programming. RL enables an exploration-exploitation trade-off, that balances between picking the known most-rewarding action and searching for a potentially better action. An optimal policy is thereby learned by a RL agent via iterative interactions (take an action and receive feedback) with the underlying dynamics, without having to know the full information of the underlying dynamics’ model, such as the coefficients of a diffusion process. While traditional RL (Sutton & Barto, 2018) is built on discrete-time Markov Decision Processes, H. Wang et al. (2020) studied continuous-time RL via a stochastic control formulation, often known as the *exploratory control problem*. H. Wang et al. (2020) extended the control from a deterministic function of time and state to a probability distribution, enabling policy exploration by randomly sampling actions from the distribution-valued policy. A series of subsequent works further fill the gaps between RL and stochastic control theory, providing theoretical foundation to some well-established RL methods — policy evaluation methods (Jia & Zhou, 2022a), policy gradient methods (Jia & Zhou, 2022b) and the Q-learning method (Jia & Zhou, 2023). Notably, the idea of exploratory control has shown its success in applications to classical finance problems, such as (but not limited to) the mean-variance portfolio optimization problem (H. Wang & Zhou, 2020; Dai et al., 2023; Chen et al., 2025b) and the optimal exercise of American options problem (Dai et al., 2024; Dong, 2024).

Dai et al. (2024) studied the optimal stopping problem under the exploratory control framework, solving for a Bernoulli distribution-valued stopping policy that represents the probability of stopping and continuing. We note that Dai et al. (2024) transformed the stopping problem into a time-consistent control problem; their optimal stopping policy is a precommitted strategy that is globally optimal at the initial time. Other recent and relevant work on exploratory optimal stopping includes (Dianetti et al., 2024; Dong, 2024). Dong (2024) directly randomized the stopping times with a Cox process and used an unnormalized negentropy as the regularizer. Dianetti et al. (2024) studied an infinite horizon stopping problem and formulated it as a singular control problem. In contrast, Dai

et al. (2024) considered a finite planning horizon and reshaped the stopping problem into a regular stochastic control problem via *penalty approximation*, effectively transforming the free-boundary PDE into a penalized PDE that can be considered as the HJB of a standard control problem. Also considering a finite horizon, this chapter adopts the idea of *penalty approximation*, as once we are in the standard control setup, all the already established exploratory control theories and RL techniques (Jia & Zhou, 2022a; H. Wang et al., 2020) are immediately applicable.

While this paper stands in the same line with (Dai et al., 2024), it extends (Dai et al., 2024) by accounting for a general (not necessarily exponential) discounting function, which introduces time-inconsistency to the problem. Time-inconsistency arises when agents' preferences vary over time, which can be modeled by a suitably chosen non-exponential discounting function. As a result, a stopping policy that is optimal right now can be suboptimal when we re-evaluate it at a later time. This chapter addresses the time-inconsistency via an *equilibrium approach* (Björk et al., 2021), deriving the Extended Hamilton-Jacobi-Bellman (EHJB) equations as a system of PDEs and solving for a pair of functions consisting of the value function and an auxiliary value function. The resulting policy is a game-theoretical equilibrium of an agent competing against her future selves, such that any deviation from this policy will make the agent worse-off. Time-inconsistent exploratory control problems have been studied in the recent mathematical finance literature. Dai et al. (2023) and Chen et al. (2025c) revisited the mean-variance portfolio optimization problem and accounted for the time-inconsistency arising from the variance term (a non-linear function of an expectation) in the objective function. Chen et al. (2025a) studied Merton's investment-consumption problem and modeled time-varying preferences with a non-exponential discounting function.

We highlight three main contributions of this chapter. First, we study a time-inconsistent exploratory stopping problem with the incorporation of a general, potentially non-exponential, discounting function. With entropy regularization, we derive the EHJB equations and solve for an equilibrium Bernoulli stopping policy in a semi-analytical form. Second, we compare exponential and non-exponential discounting, under the classical (non-exploratory) optimal stopping and the exploratory stopping frameworks. We establish two interesting results: with non-exponential discounting, (i) the stopping region is smaller under the classical framework and (ii) the Bernoulli stopping policy (as the probability of stopping) is lower under the exploratory framework, compared to the corresponding exponential discounting counterpart that has the same mean discount rate as the non-exponential discounting. This reflects the more patient behavior of a non-exponential discounting agent that prefers to wait for a potentially better future payoff, as opposed to an otherwise equivalent exponential discounting agent that suffers from a higher cost of waiting. Third, we leverage the

martingale properties from our theoretical results and develop a RL algorithm. Our algorithm trains a neural network against a martingale loss (derived following the arguments in (Jia & Zhou, 2022a)), iteratively learning the value function and the stopping policy without the knowledge of the underlying model dynamics. Meanwhile, we also numerically solve the value function and the auxiliary value function via the finite difference method on a 3-dimensional time-state-time grid, which relies on full access to the underlying model dynamics. Using the numerically-solved value functions as the benchmark, our RL algorithm successfully constructs the correct free boundary in a simulation study, effectively learning an equilibrium exploratory stopping policy.

The remainder of this chapter is structured as follows. We review the major results of (Dai et al., 2024) on the exploratory optimal stopping problem in Section 5.2, and provide an example of the time-inconsistency issue in the classical optimal stopping problem. Then, in Section 5.3, we extend the exploratory optimal stopping problem to non-exponential discounting and derive an equilibrium stopping policy. We also compare the difference in the results (i.e., stopping regions and stopping policies) under exponential and non-exponential discounting. In Section 5.4, we introduce our two algorithms, a RL algorithm and a numerical algorithm. Based on an American put option example, we implement our algorithms on simulated data and present the results in Section 5.5. Finally, we conclude in Section 5.6.

5.2 Optimal Stopping with Precommitted Strategy

In this section, we review the classical optimal stopping problem and its exploratory extension studied by (Dai et al., 2024), which serves as the theoretical foundation of this chapter. Dai et al. (2024) reformulated the optimal stopping problem as a classical stochastic control problem via a *penalty approximation*, which enables them to further extend the problem with an exploratory control under an RL framework. The main result of (Dai et al., 2024), whose notation we adopt in this chapter, focuses on the time-consistent optimal stopping problem in a finite planning horizon, which yields a precommitted optimal stopping policy. We note, however, that the classical optimal stopping problem can potentially be time-inconsistent, in which case a precommitted strategy no longer exists. We use an example to demonstrate this.

5.2.1 Classical Optimal Stopping

Consider a filtered probability space defined on a finite horizon $[0, T]$, $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})_{t \in [0, T]}$, which supports a Brownian motion $\{W_t\}_{t \in [0, T]}$. Over the planning horizon $[0, T]$, an agent decides when to stop, and collects the terminal payoff at that stopping time. The optimal stopping problem is

$$\sup_{\tau \in \mathcal{T}_{[0, T]}} \mathbb{E} [g(\tau, X_\tau)], \quad (5.1)$$

where $g : [0, T] \times \mathbb{R} \mapsto \mathbb{R}$ is the terminal payoff and $\tau \in \mathcal{T}_{[0, T]}$ is the set of all stopping times in $[0, T]$ ³. We refer to Problem (5.1) as the *classical optimal stopping problem*.

The state process $X := \{X_t\}_{t \in [0, T]}$ is driven by

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \quad (5.2)$$

where the *drift*, $b : [0, T] \times \mathbb{R} \mapsto \mathbb{R}$, and the *diffusion* $\sigma : [0, T] \times \mathbb{R} \mapsto \mathbb{R}$ are deterministic functions of time and state. Note that the state process is adapted to the filtration but is not controlled. We use the following assumption to ensure a unique solution of the SDE (5.2) exists.

Assumption 5.2.1 *b and σ are continuous in time t and Lipschitz continuous in the state x , with linear growth in x , i.e., there exist constants $C_1, C_2 > 0$ such that for all $t \in [0, T]$ and $x, y \in \mathbb{R}$,*

$$\begin{aligned} (i) \quad & |b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq C_1|x - y| \\ (ii) \quad & |b(t, x)| + |\sigma(t, x)| \leq C_2(1 + |x|). \end{aligned}$$

The *infinitesimal generator* $\mathcal{L}$ acts on a function $\psi \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R})$ by

$$\mathcal{L}\psi = \psi_t + b(t, x)\psi_x + \frac{1}{2}\sigma^2(t, x)\psi_{xx}. \quad (5.3)$$

The variational inequality of the classical optimal stopping problem, i.e., Problem (5.1), is

$$\sup\{\mathcal{L}V, g - V\} = 0, (t, x) \in [0, T] \times \mathbb{R}, \quad (5.4)$$

with terminal condition $V(T, x) = g(T, x)$.

³We implicitly assume that an agent is forced to stop the game by time T and receive the stopping payoff $g(T, X_T)$, if they have not stopped over $[0, T]$.

Dai et al. (2024) considered a penalty approximation of (5.4)

$$\mathcal{L}V + K(g - V)^+ = 0, \quad (5.5)$$

with terminal condition $V(T, x) = g(T, x)$ and $K > 0$. According to Lemma 2 of (Dai et al., 2024), the solution of the penalty-approximated (5.5) converges to the solution of the original (5.4), as $K \rightarrow \infty$. Moreover, (5.5) is equivalent to the following PDE

$$\sup_{u \in \{0,1\}} \{\mathcal{L}V - KuV + Kg(t, x)u\} = 0. \quad (5.6)$$

Here, we introduce a binary classical control, $u := \{u_t\}_{t \in [0, T]}$, representing continuing the game when $u_t = 0$ and exiting the game when $u_t = 1$.

By the Verification Theorem (Theorem 2.2 in (Dai et al., 2024)), the PDE (5.6) is the HJB equation of the stochastic control problem

$$V(t, x) = \sup_{u \in \{0,1\}} J^u(t, x), \quad (5.7)$$

where

$$J^u(t, x) := \mathbb{E}_{t,x} \left[\int_t^T K e^{-K \int_t^s u_l dl} g(s, X_s) u_s ds + e^{-K \int_t^T u_l dl} g(T, X_T) \right], \quad (5.8)$$

is the objective function and $\mathbb{E}_{t,x}[\cdot] := \mathbb{E}[\cdot | X_t = x]$. Furthermore, by Lemma 2.3 of (Dai et al., 2024), the solution to the PDE (5.6) is the solution to the stochastic control problem (5.7), which transforms the optimal stopping problem into a control problem.

Note that problem (5.7) is not a standard control problem, as it controls the exponential discounting factor in (5.6) rather than the state process X . So, Dai et al. (2024) also defined a controlled state process, $R^u := \{R_t^u\}_{t \in [0, T]}$, driven by

$$dR_s^u = -KR_s^u u_s ds \quad (5.9)$$

with $R_0^u = r_0$, where the superscript u represents that R^u is controlled by u . It is clear that $R_s^u = R_t^u e^{-K \int_t^s u_l dl}$ for $s \geq t$. Consider a new stochastic control problem

$$\bar{V}(t, x, r) = \sup_{u \in \{0,1\}} \bar{J}^u(t, x, r), \quad (5.10)$$

where

$$\bar{J}^u(t, x, r) = \mathbb{E}_{t,x,r} \left[\int_t^T K R_s^u g(s, X_s) u_s ds + R_T^u g(T, X_T) \right], \quad (5.11)$$

and $\mathbb{E}_{t,x,r}[\cdot] := \mathbb{E}[\cdot | X_t = x, R_t^u = r]$. Clearly, $J^u(t, x) = \bar{J}^u(t, x, 1)$ and $V(t, x) = \bar{V}(t, x, 1)$. Thus, Problem (5.7) is equivalent to Problem (5.10) when R^u starts from 1. The corresponding HJB equation is, for $(t, x, r) \in [0, T) \times \mathbb{R} \times (0, \infty)$,

$$\sup_{u \in \{0,1\}} \left\{ \mathcal{L}\bar{V} - Kru \frac{\partial \bar{V}}{\partial r} + Krg(t, x)u \right\} = 0, \quad (5.12)$$

with terminal and boundary conditions $\bar{V}(T, x, r) = rg(T, x)$ and $\bar{V}(t, x, 0) = 0$.

5.2.2 Exploratory Extension to Classical Optimal Stopping

Dai et al. (2024) considered an exploratory extension to problem (5.10). Let $\pi_{t,X_t} : [0, T] \times \mathbb{R} \mapsto [0, 1]$ be a mapping from a time-state-pair, (t, X_t) , to a Bernoulli distribution over the control space $\{0, 1\}$. For ease of notation, we hereafter write π_t as shorthand for π_{t,X_t} unless otherwise clarified. The exploratory control, $\pi := \{\pi_t\}_{t \in [0, T]}$, is a Bernoulli distribution-valued stochastic process adapted to the filtration, defined on the finite control space $\{0, 1\}$, where $\pi_t := \mathbb{P}(u_t = 1)$ is the probability of exiting the game and $1 - \pi_t = \mathbb{P}(u_t = 0)$ is the probability of continuing the game. The controlled state process extends from (5.9) to

$$dR_s^\pi = -KR_s^\pi \pi_s ds \quad (5.13)$$

hence $R_s^\pi = R_t^\pi e^{-K \int_t^s \pi_l dl}$ for $s \geq t$. The state process X (5.2) is unchanged as it is uncontrolled. The objective function, with an inclusion of entropy regularization, extends from (5.11) to

$$\bar{J}^\pi(t, x, r) = \mathbb{E}_{t,x,r} \left[\int_t^T R_s^\pi \left[Kg(s, X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right] ds + R_T^\pi g(T, X_T) \right], \quad (5.14)$$

for $(t, x, r) \in [0, T] \times \mathbb{R} \times \mathbb{R}_+$, where $\lambda > 0$ is the exploration weight. Then, the entropy regularized exploratory extension to Problem (5.10) is

$$\bar{V}(t, x, r) = \sup_{\pi \in [0,1]} \bar{J}^\pi(t, x, r). \quad (5.15)$$

For $(t, x, r) \in [0, T) \times \mathbb{R} \times \mathbb{R}_+$, the HJB becomes

$$\mathcal{L}\bar{V}(t, x, r) + r \cdot \sup_{\pi \in [0,1]} \{-K\pi \bar{V}_r(t, x, r) + Kg(t, x)\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} = 0, \quad (5.16)$$

where $\bar{V}_r := \frac{\partial}{\partial r} \bar{V}$. Note that since X_t is not controlled, the infinitesimal generator $\mathcal{L}$ is the same as the non-exploratory case, defined in (5.3). Solving the supremum part of (5.16) with the first order condition yields

$$\bar{\pi}^* = \frac{1}{1 + \exp(\frac{1}{\lambda} K(\bar{V}_r - g))}. \quad (5.17)$$

Dai et al. (2024) showed that (5.17) is equivalent to

$$\pi^* = \frac{1}{1 + \exp(\frac{1}{\lambda} K(V - g))}. \quad (5.18)$$

Here, V is the solution to the entropy regularized exploratory extension to Problem (5.7). That is,

$$V := V(t, x) = \sup_{\pi \in [0,1]} J^\pi(t, x), \quad (5.19)$$

where

$$\begin{aligned} J^\pi(t, x) = \mathbb{E}_{t,x} \left[\int_t^T e^{-K \int_t^s \pi_l dl} \left(K g(s, X_s^\pi) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right. \\ \left. + e^{-K \int_t^T \pi_l dl} g(T, X_T^\pi) \right] \end{aligned} \quad (5.20)$$

comes from substituting R_s^π into (5.14) and the corresponding HJB equation is given by

$$\mathcal{L}V(t, x) + \sup_{\pi \in [0,1]} \{-K\pi V(t, x) + K g(t, x)\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} = 0, \quad (5.21)$$

for $(t, x) \in [0, T] \times \mathbb{R}$. The following relationships hold between the results of Problem (5.15) and Problem (5.19)

$$\begin{aligned} J^\pi(t, x) &= \frac{1}{r} \bar{J}^\pi(t, x, r) \\ V(t, x) &= \frac{1}{r} \bar{V}(t, x, r) = \bar{V}(t, x, 1) \\ \bar{V}_r(t, x, r) &= V(t, x), \end{aligned}$$

for $(t, x, r) \in [0, T] \times \mathbb{R} \times \mathbb{R}_+$.

5.2.3 Time Inconsistent Issue in Classical Optimal Stopping

A time-consistent problem requires the solved strategy to be optimal at any point of time in the planning horizon, i.e., the optimal control at time $t_1 \in [0, T]$ is also optimal at any other time $t_2 \in [0, T]$, regardless of $t_1 > t_2$ or $t_1 < t_2$ (Björk et al., 2021). Therefore, a precommitted strategy that is globally optimal over the entire planning horizon can be solved at the initial time $t = 0$. While Dai et al. (2024) solved for a globally optimal *precommitted* strategy, the classical optimal stopping problem (Problem (5.1)) can be time-inconsistent so that a precommitted strategy that is optimal at time $t_1 \in [0, T]$ is not necessarily optimal at a different time $t_2 \in [0, T]$ with $t_1 \neq t_2$. We use the following example to illustrate this issue.

Consider a special case of Problem (5.1) with $g(t, x) = \beta_t \tilde{g}(x)$, for some discounting function $\beta_t : [0, T] \mapsto [0, 1]$ satisfying $\beta_0 = 1$ and the terminal payoff function $\tilde{g} : \mathbb{R} \mapsto \mathbb{R}$. Suppose the terminal payoff function is a standard put option payoff

$$\tilde{g}(x) = (K - x)^+,$$

where $K = \$40$ is the strike price. Further suppose that the state process (5.2) is a stock price dynamic driven by a Geometric Brownian Motion

$$dX_t = \mu X_t dt + \sigma X_t dW_t,$$

where $(\mu, \sigma) = (0.06, 0.4)$ are the mean and volatility of the stock returns and the stock price is initialized at $X_0 = \$40$. Consider an exponential discounting function, $\beta_t^{exp} = e^{-\rho t}$ with $\rho = 0.04$, and a gamma (non-exponential) discounting function, $\beta_t^{nonexp} = (1 + \alpha t)^{\frac{\beta}{\alpha}}$ with $(\alpha, \beta) = (0.0225, 0.04)$.⁴ We follow Section 5.2.1 and numerically solve the HJB (5.6) with $g(t, x)$ replaced by $\beta_t \tilde{g}(x)$ via the finite difference method, at two selected initial time points, $t_0 = 0$ and $t_0 = 0.2$. Then, we calibrate the precommitted free boundaries for exercise of the put options, at the two initial times and with the two discounting functions. As shown in Figure 5.1, the free boundaries illustrate the precommitted stopping policies, as an investor will exercise (stop holding) the put option whenever the stock price drops below the free boundaries.

Note that this problem is time-consistent with exponential discounting and is time-inconsistent with non-exponential discounting (Björk et al., 2021). The free boundaries collapse on the left of Figure 5.1, indicating that the precommitted stopping strategy is

⁴Gamma discounting is a typical non-exponential discounting function, described in (Weitzman, 2001), from which the parameters (α, β) are selected. We will continue using these parameter settings in Section 5.5, with detailed explanations provided.

globally optimal and unchanged at different times in the time-consistent setting with exponential discounting. In contrast, the free boundaries on the right of Figure 5.1 are distinct from each other, implying that the optimal precommitted stopping strategies vary at different times in the time-inconsistent setting with non-exponential discounting. This is a typical case of time-inconsistent stopping problem, where the globally optimal precommitted strategy does not exist. Henceforth, we redirect ourselves to an equilibrium approach (Björk et al., 2021), which solves for a optimal strategy that is locally optimal at each time point over the planning horizon.

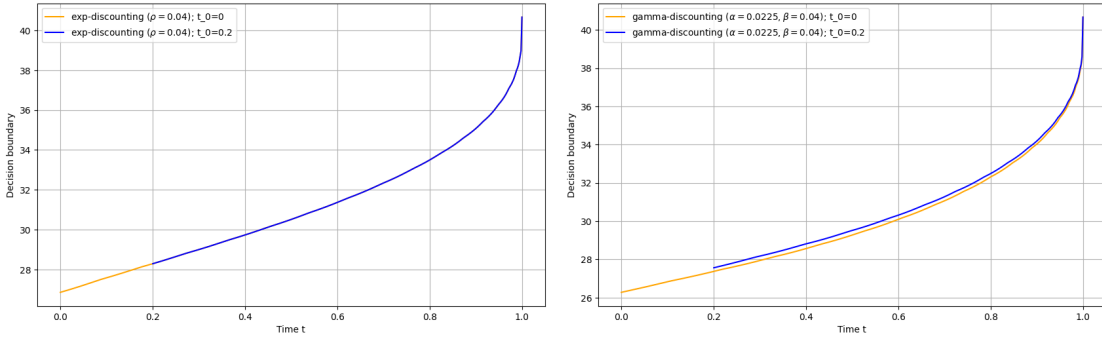


Figure 5.1: Free boundaries from different initial times $t_0 = 0$ vs $t_0 = 0.2$ for exponential (left) and non-exponential (right) discounting.

From here on, we derive theoretical results based on the original stochastic control problem formulation (without the controlled process R), as this approach is more direct. Nevertheless, we develop an RL algorithm under the new stochastic control problem formulation (with the controlled process R), as it is easier for implementation.

5.3 Equilibrium Exploratory Optimal Stopping with Non-Exponential Discounting

In this section, we focus on a special case of Problem (5.1) that we described in Section 5.2.3. We introduce a discounting function for the aforementioned exploratory optimal stopping problem in Section 5.2.2 and study some formal results. While Dai et al. (2024) focused on the time-consistent case, we extend it with a general discounting function, not necessarily exponential, and derive the optimal stopping strategy. Accounting for a possibly

non-exponential discounting function transitions this problem to a time-inconsistent problem, yielding a stopping strategy that is optimal in a game-theoretical equilibrium sense. Beyond that, we also compare the difference in the optimal stopping strategies, including the stopping regions and stopping policies, implied by exponential and non-exponential discounting.

5.3.1 Equilibrium Policy and Verification Theorem

Suppose $\beta_t : [0, T] \mapsto [0, 1]$ is a non-exponential differentiable discounting function satisfying $\beta_0 = 1$. We consider a special case of Problem (5.1), referred to as the *discounted optimal stopping problem*,

$$\sup_{\tau \in \mathcal{T}_{[0, T]}} \mathbb{E} [\beta_\tau g(X_\tau)], \quad (5.22)$$

where $g : \mathbb{R} \mapsto \mathbb{R}$ is the terminal payoff function that only depends on the state. Recall from Section 5.2.3 that this problem is time-inconsistent due to the inclusion of a non-exponential discounting function. Hence, a globally optimal precommitted strategy is unavailable and we have to solve for a strategy that is locally optimal at each time point. We can interpret this as an agent's preferences changing over time, so that a policy that the agent is committed to right now can be changed at a later time, when the policy is re-evaluated. Henceforth, an agent competes with her future selves and compromises to a game-theoretical equilibrium, such that any deviation from it at any time will make the agent worse-off. The resulting policy is therefore referred to as an *equilibrium policy*, which we formally define as follows.

Definition 5.3.1 (Equilibrium Policy) *Let $\pi^* \in [0, 1]$ be a stopping policy. For any $t \in [0, T]$, consider a perturbed policy $\pi^{\varepsilon, v} := \{\pi_s^{\varepsilon, v}\}_{s \in [t, T]}$, defined as*

$$\pi_s^{\varepsilon, v} = \begin{cases} v, & s \in [t, t + \varepsilon), \\ \pi_s^*, & s \in [t + \varepsilon, T], \end{cases}$$

where $\varepsilon > 0$, and $v \in [0, 1]$ is a constant stopping probability. Suppose that for any $(t, x) \in [0, T] \times \mathbb{R}$ and any such perturbed strategy $\pi^{\varepsilon, v}$, we have

$$\limsup_{\varepsilon \downarrow 0} \frac{J_{ne}^{\pi^{\varepsilon, v}}(t, x) - J_{ne}^{\pi^*}(t, x)}{\varepsilon} \leq 0.$$

Then π^* is an equilibrium stopping policy to Problem (5.24) and $J^{\pi^*}(t, x)$ is the corresponding value function.

Denote by J_{ne}^π the entropy-regularized objective function with non-exponential discounting, which is extended from (5.20) to

$$J_{ne}^\pi(t, x) = \mathbb{E}_{t,x} \left[\int_t^T e^{-K \int_t^s \pi_l dl} \beta_{s-t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds + e^{-K \int_t^T \pi_l dl} \beta_{T-t} g(X_T) \right]. \quad (5.23)$$

The exploratory control problem becomes

$$V_{ne}(t, x) = \sup_{\pi \in [0,1]} J_{ne}^\pi(t, x). \quad (5.24)$$

With the subscript “ne”, V_{ne} represents the value function with non-exponential discounting. We note that Problem (5.24) is time-inconsistent, due to the presence of a non-exponential discounting function. Define an *auxiliary objective function* for a given policy π and $(t, x, s) \in [0, T] \times \mathbb{R} \times [0, t]$,

$$f^\pi(t, x, s) = \mathbb{E}_{t,x} \left[\int_t^T e^{-K \int_t^v \pi_l dl} \beta_{v-s} \left(Kg(X_v) \pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right) dv + e^{-K \int_t^T \pi_l dl} \beta_{T-s} g(X_T^\pi) \right]. \quad (5.25)$$

Comparing (5.23) to (5.25), the objective function J_{ne}^π is discounted to a given time t , whereas the auxiliary objective function $f^\pi(t, x, s)$ is discounted to an even earlier time $s \leq t$. Clearly, $f^\pi(t, x, t) = J_{ne}^\pi(t, x)$, for all $(t, x) \in [0, T] \times \mathbb{R}$. Furthermore, denote by $f(t, x, s) = f^{\pi^*}(t, x, s)$ the *auxiliary value function*, where π^* is the maximizer of Problem (5.24).

Following the equilibrium approach of time-inconsistent control problems in (Björk et al., 2021), we consider the EHJB equations for Problem (5.24):

$$\mathcal{L}V + \sup_{\pi \in [0,1]} \{-K\pi V + Kg\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} - \frac{\partial f}{\partial s}(t, x, t) = 0 \quad (5.26)$$

$$\mathcal{L}f(t, x, s) - K\pi^* f(t, x, s) + \beta_{t-s}[Kg(x)\pi^* - \lambda(\pi^* \log \pi^* + (1 - \pi^*) \log(1 - \pi^*))] = 0 \quad (5.27)$$

with terminal conditions $V(T, x) = g(x)$ and $f(T, x, s) = \beta_{T-s}g(x)$. Here, $\frac{\partial f}{\partial s}$ is the first order partial derivative of f with respect to its third argument. The derivations of (5.26) and (5.27) are given in Appendix A.4.1.

Remark 5.3.1 *If we instead consider a classical (non-exploratory) policy, $u := \{u_t\}_{t \in [0, T]}$ as introduced in Section 5.2, then the optimal stopping problem with discounting becomes Problem (5.7) with J^u replaced by*

$$J_{ne}^u(t, x) := \mathbb{E}_{t,x} \left[\int_t^T K e^{-K \int_t^s u_l dl} \beta_{s-t} g(X_s) u_s ds + e^{-K \int_t^T u_l dl} \beta_{T-t} g(X_T) \right], \quad (5.28)$$

for $(t, x) \in [0, T] \times \mathbb{R}$. Define the auxiliary objective function as

$$f_{ne}^u(t, x, s) := \mathbb{E}_{t,x} \left[\int_t^T K e^{-K \int_t^v u_l dl} \beta_{v-s} g(X_v) u_v dv + e^{-K \int_t^T u_l dl} \beta_{T-s} g(X_T) \right], \quad (5.29)$$

for $(t, x, s) \in [0, T] \times \mathbb{R} \times [0, t]$. The EHJB system with penalty approximation becomes

$$\mathcal{L}V + K(g - V)^+ - \frac{\partial f}{\partial s}(t, x, t) = 0 \quad (5.30)$$

$$\mathcal{L}f(t, x, s) + K(\beta_{t-s}g(x) - f(t, x, s)) \mathbb{I}_{V(t,x)=g(x)} = 0, \quad (5.31)$$

with terminal condition $V(T, x) = g(x)$ and $f(T, x, s) = \beta_{T-s}g(x)$.

Clearly, solving the supremum part of (5.26) yields

$$\pi_{ne}^* = \frac{1}{1 + \exp(\frac{1}{\lambda} K(V_{ne} - g))}, \quad (5.32)$$

which takes exactly the same form as the precommitted stopping policy (5.18). We remark here that the precommitted stopping policy π^* (5.18) is attained by solving a time-consistent stochastic control problem, hence it is globally optimal over the planning horizon. In contrast, the non-exponential discounted stopping policy π_{ne}^* (5.32) is a solution⁵ to a time-inconsistent stochastic control problem, which is locally optimal in a game-theoretical equilibrium sense.

We summarize our theoretical results in the following theorem. The proof is presented in Appendix A.4.2.

⁵The uniqueness of the solution to a time-inconsistent control problem is an open question (Dai et al., 2023).

Theorem 5.3.1 (Verification Theorem) *Suppose that there exist $V(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R})$ and $f(t, x, s) \in \mathbb{C}^{1,2,1}([0, T] \times \mathbb{R} \times [0, t])$ satisfying (5.26) and (5.27), with terminal conditions $V(T, x) = g(x)$ and $f(T, x, s) = \beta_{T-s}g(x)$. Then, π_{ne}^* , given by (5.32), is an equilibrium solution to Problem (5.24). Moreover, $V(t, x) = J_{ne}^{\pi^*}(t, x)$ with J_{ne}^{π} given by (5.23) and $f(t, x, s) = f^{\pi^*}(t, x, s)$ with $f^{\pi}(t, x, s)$ given by (5.25).*

As also remarked by ((Dai et al., 2024)), Problem 5.24 is not a conventional stochastic control problem, due to the presence of the control in the discounting factors. Luckily, the same result follows if we adopt the notation of the controlled process R from (5.13), which we repeat below for readers' convenience

$$dR_s^\pi = -K R_s^\pi \pi_s ds.$$

Consider a new stochastic control problem

$$\bar{V}_{ne}(t, x, r) = \sup_{\pi \in [0,1]} \bar{J}_{ne}^\pi(t, x, r), \quad (5.33)$$

where the new entropy-regularized objective function

$$\begin{aligned} \bar{J}_{ne}^\pi(t, x, r) = \mathbb{E}_{t,x,r} \left[\int_t^T R_s^\pi \beta_{s-t} \left(K g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right. \\ \left. + R_T^\pi \beta_{T-t} g(X_T) \right], \end{aligned} \quad (5.34)$$

for $(t, x, r) \in [0, T] \times \mathbb{R} \times \mathbb{R}_+$, is transitioned from (5.23). Similarly, the auxiliary objective function transitions from (5.25) to

$$\begin{aligned} \bar{f}^\pi(t, x, s, r) = \mathbb{E}_{t,x} \left[\int_t^T R_v^\pi \beta_{v-s} \left(K g(X_v) \pi_v - \lambda (\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right) dv \right. \\ \left. + R_T^\pi \beta_{T-s} g(X_T^\pi) \right], \end{aligned} \quad (5.35)$$

for $(t, x, s, r) \in [0, T] \times \mathbb{R} \times [0, t] \times \mathbb{R}_+$. Denote by $\bar{f}(t, x, s, r) := \bar{f}^{\pi^*}(t, x, s, r)$ the *auxiliary value function*, for $(t, x, s, r) \in [0, T] \times \mathbb{R} \times [0, t] \times \mathbb{R}_+$. The EHJB equations are updated

from (5.26) and (5.27) to

$$\mathcal{L}\bar{V} + r \cdot \sup_{\pi \in [0,1]} \{-K\pi\bar{V}_r + Kg\pi - \lambda(\pi \log \pi + (1-\pi) \log(1-\pi))\} - \frac{\partial \bar{f}}{\partial s}(t, x, t, r) = 0 \quad (5.36)$$

$$\mathcal{L}\bar{f}(t, x, s, r) - r \cdot \left\{ K\pi^*\bar{f}_r(t, x, s, r) + \beta_{t-s}[Kg(x)\pi^* - \lambda(\pi^* \log \pi^* + (1-\pi^*) \log(1-\pi^*))] \right\} = 0, \quad (5.37)$$

where $\bar{V}_r := \frac{\partial \bar{V}}{\partial r}(t, x, r)$ and $\bar{f}_r(t, x, s, r) = \frac{\partial \bar{f}}{\partial r}(t, x, s, r)$. One can check that

$$\bar{\pi}_{ne}^* = \frac{1}{1 + \exp(\frac{1}{\lambda}K(\frac{\partial}{\partial r}\bar{V}_{ne} - g))} \quad (5.38)$$

solves the supremum part of (5.36), which matches the form of the optimal precommitted stopping policy in the time-consistent case (5.17). Consequently, the following relationships hold between the results of Problem (5.24) and Problem (5.33)

$$\begin{aligned} J_{ne}^\pi(t, x) &= \frac{1}{r} \bar{J}_{ne}^\pi(t, x, r) \\ V_{ne}(t, x) &= \frac{1}{r} \bar{V}_{ne}(t, x, r) = \bar{V}_{ne}(t, x, 1) \\ f^\pi(t, x, s) &= \frac{1}{r} \bar{f}^\pi(t, x, s, r) \\ f(t, x, s) &= \frac{1}{r} \bar{f}(t, x, s, r) = \bar{f}(t, x, s, 1) \\ \frac{\partial}{\partial r} \bar{V}_{ne}(t, x, r) &= V_{ne}(t, x), \end{aligned}$$

for $(t, x, s, r) \in [0, T] \times \mathbb{R} \times [0, t] \times \mathbb{R}_+$. Clearly, we have the following lemma.

Lemma 5.3.1.1 *Suppose that there exist $\bar{V}_{ne}(t, x, r) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R} \times \mathbb{R}_+)$ and $\bar{f}(t, x, s, r) \in \mathbb{C}^{1,2,1}([0, T] \times \mathbb{R} \times [0, t] \times \mathbb{R}_+)$ satisfying (5.36) and (5.37), with terminal conditions $\bar{V}_{ne}(T, x, r) = rg(x)$ and $\bar{f}(T, x, s, r) = \beta_{t-s}rg(x)$. Then, $\bar{\pi}_{ne}^*$, given by (5.38), is an equilibrium solution to Problem (5.33). Moreover, $\bar{V}_{ne}(t, x, r) = \bar{J}_{ne}^{\pi^*}(t, x, r)$ with $\bar{J}_{ne}^{\pi^*}$ given by (5.34) and $\bar{f}(t, x, s, r) = \bar{f}^{\pi^*}(t, x, s, r)$ with $\bar{f}^{\pi^*}$ given by (5.35).*

The proof is similar to that of Theorem 5.3.1, hence omitted.

5.3.2 Comparison Between Exponential and Non-Exponential Discounting

We devote the rest of this section to a comparison of stopping policies under exponential and non-exponential discounting, with classical policies and exploratory policies. We suppose a non-exponential discounting function is a *weighted discounting function*, which, according to Definition 1 in (Ebert et al., 2020), can be expressed as an expectation of exponential discounting with random discount rates following a probability distribution F , i.e.,

$$\beta_t^{\text{nonexp}} = \int_0^\infty e^{-rt} dF(r). \quad (5.39)$$

So, $(\beta_t^{\text{nonexp}})' = -\int_0^\infty r e^{-rt} dF(r) < 0$, i.e., β_t^{nonexp} is decreasing and differentiable. Denote by

$$\beta_t^{\text{exp}} = e^{-\rho t} \quad (5.40)$$

the exponential discounting function with $\rho > 0$, a positive constant discount rate. For a fair comparison between β_t^{nonexp} and β_t^{exp} , we further require the mean discount rates for exponential and non-exponential discounting are equal, i.e.,

$$\int_0^\infty r dF(r) = \rho. \quad (5.41)$$

Recall the discounted optimal stopping problem (5.22), with β_t replaced with β_t^{exp} and β_t^{nonexp} for the exponential and non-exponential discounting cases respectively. With a classical stopping policy, the HJB equations are

$$\text{exponential discounting: } \sup\{\mathcal{L}V_{\text{exp}} - \rho V_{\text{exp}}, g - V_{\text{exp}}\} = 0 \quad (5.42)$$

$$\text{non-exponential discounting: } \sup\{\mathcal{L}V_{\text{ne}} - \frac{\partial f}{\partial s}(t, x, t), g - V_{\text{ne}}\} = 0, \quad (5.43)$$

with terminal conditions $V_{\text{exp}}(T, x) = g(x)$ and $V_{\text{ne}}(T, x) = g(x)$. Here, for $(t, x, s) \in [0, T] \times \mathbb{R} \times [0, t]$, V_{exp} and V_{ne} are respectively the value functions under exponential and non-exponential discounting. $f(t, x, s)$ is the auxiliary value function that is only different from $V_{\text{ne}}(t, x)$ by the time to which it is discounted. While $V_{\text{ne}}(t, x)$ is discounted to t , $f(t, x, s)$ is discounted to $s \leq t$, with $f(t, x, t) = V_{\text{ne}}(t, x)$. The $\frac{\partial}{\partial s}$ operator is only applied to the discounting function in f . It is well-known that the stopping regions are

$$\text{exponential discounting: } D_{\text{exp}} = \{(t, x) \in [0, T] \times \mathbb{R} : V_{\text{exp}}(t, x) = g(x)\} \quad (5.44)$$

$$\text{non-exponential discounting: } D_{\text{ne}} = \{(t, x) \in [0, T] \times \mathbb{R} : V_{\text{ne}}(t, x) = g(x)\}. \quad (5.45)$$

Proposition 5.3.1 *With classical stopping policies, the stopping region is smaller under non-exponential discounting than exponential discounting, i.e., $D_{ne} \subseteq D_{exp}$, under condition (5.41).*

The proof is provided in Appendix A.4.3. With classical policies, an agent immediately stops when the value of continuing the game (V_{exp} or V_{ne}) equals the value of stopping g . As we show in the proof of Proposition 5.3.1, non-exponential discounting dominates exponential discounting, i.e., $\beta_t^{\text{nonexp}} \geq \beta_t^{\text{exp}}, \forall t \in [0, T]$, when (5.41) holds. This effectively means a non-exponential discounting agent is penalized less for waiting, thus being more patient and having higher value of continuing the game than an exponential discounting agent. As a result, a non-exponential discounting agent is more expectant for a better state x at each given time $t \in [0, T]$ and more willing to wait for a better time at each given state $x \in \mathbb{R}$, yielding a smaller stopping region than an otherwise equivalent exponential discounting agent.

On the other hand, with exploratory stopping policies and a constant $K > 0$ for penalty approximation, (5.32) implies that the optimal stopping policy for exponential discounting and an equilibrium stopping policy for non-exponential discounting⁶ are respectively given by

$$\text{exponential discounting: } \pi_{exp}^* = \frac{1}{1 + \exp(\frac{1}{\lambda}K(V_{exp} - g))}, \quad (5.46)$$

$$\text{non-exponential discounting: } \pi_{ne}^* = \frac{1}{1 + \exp(\frac{1}{\lambda}K(V_{ne} - g))}. \quad (5.47)$$

One can easily check that π_{exp}^* is the optimal solution to Problem (5.22) if we set $\beta_t = \beta_t^{\text{exp}}$. In that case, the HJB equation becomes

$$\mathcal{L}V + \sup_{\pi \in [0,1]} \{-K\pi V + Kg\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} - \rho V = 0. \quad (5.48)$$

Proposition 5.3.2 *The equilibrium stopping policy under non-exponential discounting is lower than the optimal stopping policy under exponential discounting, i.e., $\pi_{ne}^* \leq \pi_{exp}^*$, under condition (5.41).*

⁶We herein distinguish the terms of the solved stopping policy for exponential discounting, π_{exp}^* , and non-exponential discounting, π_{ne}^* , as they are “optimal” in different senses. Under exponential discounting, π_{exp}^* is the solution to a time-consistent control problem, hence it is globally optimal. In contrast, π_{ne}^* is an equilibrium solution to a time-inconsistent control problem due to the presence of the non-exponential discounting, which is only locally optimal.

A proof is presented in Appendix A.4.4. We remark that with exploratory stopping policies, an agent does not necessarily stop when the value of continuation, V_{exp}^* or V_{ne}^* , equals (or is even lower than) the value of stopping g . Rather, an agent decides whether to stop with a probability equal to π_{exp}^* or π_{ne}^* , which is 50% when V_{exp} or V_{ne} equals g . That is, the probability of stopping is 50% on the stopping region of the corresponding classical stopping policy under exponential or non-exponential discounting, D_{exp} or D_{ne} . Proposition 5.3.2 effectively states that the probability of stopping under non-exponential discounting is lower than that under exponential discounting, given that the mean discount rates of the exponential and non-exponential discounting functions are equal and all other problem variables are the same. The implication is similar to that of the aforementioned classical policies — a non-exponential discounting agent has a lower chance of stopping the game at each time $t \in [0, T]$ and state $x \in \mathbb{R}$, being more patient (and greedy) for a better payoff in the future, compared to an otherwise equal exponential discounting agent.

5.4 RL and Finite Difference Algorithms

In this section, we design two algorithms, a RL algorithm and a finite difference algorithm, to learn the value function of the exploratory optimal stopping problem with discounting. We recognize that the value function does not have a closed form solution as the optimal stopping problem has a finite horizon. So, in the RL algorithm, we configure a neural network to approximate the value function, and leverage the martingale property of a specially defined process to design a tailored loss function for training. Our RL algorithm does not rely on the true parameters of the state process (5.2). On the other hand, the finite difference algorithm iteratively updates the value function on a fine grid of time and state values and relies on the knowledge of the true parameters of the state process (5.2). Nevertheless, the finite difference algorithm serves as a useful benchmark to evaluate the RL algorithm.

We start by introducing the RL algorithm. Note from (5.32) that the policy is immediately determined if the value function is known. We approximate the value function with a neural network, V^θ , parameterized by θ . Then, the parameterized stopping policy is given by

$$\pi^\theta := \frac{1}{\exp(\frac{1}{\lambda}K(V^\theta - g))}.$$

For $(t, x) \in [0, T] \times \mathbb{R}$ and a stopping policy π , define a process $M^\pi := \{M_t^\pi :=$

$M(t, X_t)\}_{t \in [0, T]}$, with

$$\begin{aligned} M_t^\pi &= \bar{J}_{ne}^\pi(t, X_t, R_t^\pi) + \int_0^t R_v^\pi \left[Kg(X_v)\pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right] dv \\ &= R_t^\pi \cdot J_{ne}^\pi(t, X_t) + \int_0^t R_v^\pi \left[Kg(X_v)\pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right] dv. \end{aligned} \quad (5.49)$$

Here, $\bar{J}_{ne}$ is given by (5.34) and $\bar{f}^\pi$ is given by (5.35). The following theorem is the foundation of our RL algorithm. A proof is provided in Appendix A.4.5.

Theorem 5.4.1 *M^π is a martingale if $\pi = \pi_{ne}^*$ given by (5.32).*

We note that if M^π is a martingale, then at any time $t \in [0, T]$, M_t^π is equivalent to the expectation of the terminal value M_T^π conditioned on all the information up to time t , namely,

$$M_t^\pi = \mathbb{E}[M_T^\pi | \mathcal{F}_t]. \quad (5.50)$$

Meanwhile, M_t^π , as a conditional expectation, minimizes the L^2 distance between M_T^π and any $\mathcal{F}_t$ -measurable random variable, that is,

$$M_t^\pi = \arg \min_{\xi \text{ is } \mathcal{F}_t\text{-measurable}} \mathbb{E}|M_T - \xi|^2, \quad (5.51)$$

for all $t \in [0, T]$. Define $M^\theta := \{M_t^\theta\}_{t \in [0, T]}$ as the θ -parameterized M^π process with

$$M_t^\theta = R_t^\theta \cdot V_{ne}^\theta(t, X_t) + \int_0^t R_v^\theta \left[Kg(X_v)\pi_v^\theta - \lambda(\pi_v^\theta \log \pi_v^\theta + (1 - \pi_v^\theta) \log(1 - \pi_v^\theta)) \right] dv, \quad (5.52)$$

where $R_t^\theta := R_t^{\pi^\theta}$ and V_{ne}^θ is the θ -parameterized value function V_{ne} with non-exponential discounting. We herein define a *martingale loss function*, following (Jia & Zhou, 2022a).

Definition 5.4.1 (Martingale Loss Function) *Denote by $ML(\theta)$ the martingale loss with parameter θ ,*

$$\begin{aligned} ML(\theta) &:= \frac{1}{2} \mathbb{E} \left[\int_0^T \left| M_T^\theta - M_t^\theta \right|^2 dt \right] \\ &= \frac{1}{2} \mathbb{E} \left[\int_0^T \left| R_T^\theta g(X_T) - R_t^\theta V_{ne}^\theta(t, X_t) \right. \right. \\ &\quad \left. \left. + \int_t^T R_v^\theta \left[Kg(X_v)\pi_v^\theta - \lambda(\pi_v^\theta \log \pi_v^\theta + (1 - \pi_v^\theta) \log(1 - \pi_v^\theta)) \right] dv \right|^2 dt \right], \end{aligned} \quad (5.53)$$

where the uncontrolled state process X is driven by (5.2).

We apply the gradient descent method to minimize $ML(\theta)$ and update θ following

$$\theta \leftarrow \theta - \eta \frac{d}{d\theta} ML(\theta),$$

for some positive learning rate η . Consider a time discretization, $0 = t_0 < \dots < t_n < \dots < t_{N_t} = T$ with a constant mesh size Δt such that $N_t = \frac{T}{\Delta t}$. We apply Euler's approximation on both the uncontrolled state process X (5.2) and the controlled state process R^π (5.13), thus for $n = 0, \dots, N_t - 1$,

$$X_{t_{n+1}} \leftarrow X_{t_n} + b(t_n, X_{t_n})\Delta t + \sigma(t_n, X_{t_n})\Delta W_{t_n} \quad (5.55)$$

$$R_{t_{n+1}}^\pi \leftarrow R_{t_n}^\pi (1 - K\pi_{t_n})\Delta t, \quad (5.56)$$

where $\Delta W_{t_n} = W_{t_{n+1}} - W_{t_n} \stackrel{iid}{\sim} N(0, \Delta t)$.

We summarize the RL algorithm in Algorithm 4.

Algorithm 4: Training with Martingale Loss

Initialize the number of epochs N_{epochs} , the planning horizon T , the mesh size of the time step Δt , the exploration parameter $\lambda > 0$, the learning rate η , the initial uncontrolled state x_0 , the initial controlled state $r_0 = 1$, the value function approximator V^θ and its parameter $\theta^{(0)}$.

Given the payoff function g and discounting function β .

Compute $N_t = \frac{T}{\Delta t}$.

for $k = 0, \dots, N_{epochs} - 1$ **do**

Fix a realized path of the Brownian motion $\{W_{t_n}^{(k)}\}_{n=0}^{N_t}$, with

$\Delta W_{t_n}^{(k)} := W_{t_{n+1}}^{(k)} - W_{t_n}^{(k)} \sim N(0, \Delta t)$.

Simulate $\{X_{t_n}^{(k)}\}_{n=0}^{N_t}$ via (5.55), with $\{\Delta W_{t_n}\}_{n=0}^{N_t}$ replaced by $\{\Delta W_{t_n}^{(k)}\}_{n=0}^{N_t}$.

for $n = 0, \dots, N_t - 1$ **do**

Compute the policy $\pi_{t_n}^{\theta^{(k)}} = \frac{1}{\exp\left(\frac{1}{\lambda} K(V^{\theta^{(k)}}(t_n, X_{t_n}) - g(X_{t_n}))\right)}$.

Compute the controlled state $R_{t_n}^{\theta^{(k)}}$ via (5.56), with π_{t_n} replaced by $\pi_{t_n}^{\theta^{(k)}}$.

end

Collect $\{\pi_{t_n}^{\theta^{(k)}}\}_{n=0}^{N_t}$ and $\{R_{t_n}^{\theta^{(k)}}\}_{n=0}^{N_t}$.

Compute $ML(\theta^{(k)})$ according to Definition 5.4.1, with time discretization and θ replaced by $\theta^{(k)}$.

Compute the gradient $\frac{d}{d\theta} ML(\theta^{(k)})$ via back propagation and update

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta \frac{d}{d\theta} ML(\theta^{(k)}).$$

end

While the aforementioned RL algorithm is designed for the non-exponential discounting optimal stopping problem, we remark that it is also applicable to the exponential discounting case. Define a process $\tilde{M}^\pi := \{\tilde{M}_t^\pi := \tilde{M}(t, X_t)\}_{t \in [0, T]}$, which is similar to (5.49) except that J_{ne}^π is replaced by

$$\begin{aligned} J_{exp}^\pi(t, x) = & \mathbb{E}_{t,x} \left[\int_t^T e^{-K \int_t^s \pi_l dl} e^{-\rho(s-t)} \left(K g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right. \\ & \left. + e^{-K \int_t^T \pi_l dl} e^{-\rho(T-t)} g(X_T) \right], \end{aligned} \quad (5.57)$$

for some exponential discounting rate $\rho > 0$. Then, we can infer the following corollary

from Theorem 5.4.1.

Corollary 5.4.1 $\tilde{M}^\pi$ is a martingale if $\pi = \pi_{exp}^*$ given by (5.47).

The proof is provided in Appendix A.4.6. Following the same procedure that we described above, we can train and learn the optimal exponential discounting stopping policy via Algorithm 4, by setting β as an exponential discounting function.

We next develop a numerical algorithm that iteratively approximates the solution of the EHJB equations with non-exponential discounting and classical stopping policy, namely (5.30) and (5.31). For demonstration purposes, we suppose the uncontrolled state process is the price of a risky asset (e.g., a stock) and we simplify it by setting $b(t, x) = \mu \cdot x$ and $\sigma(t, x) = \sigma \cdot x$. Thus (5.2) becomes

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \quad (5.58)$$

where $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ are respectively the growth rate and volatility of the stock. Note that we implicitly assume that X is non-negative, as the price cannot be negative. Replace $V(t, X_t) = v(t, Y_t)$ and $f(t, X_t, s) = h(t, Y_t, s)$, where $X_t = e^{Y_t}$ for all $t \in [0, T], s \in [0, t]$. So, by Itô's lemma,

$$dY_t = d(\log X_t) = \frac{1}{X_t} dX_t - \frac{1}{2X_t^2} d[X]_t = (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW_t.$$

Partition the 3-dimensional space of time-state-time triplets, $(t, x, s) \in [0, T] \times \mathbb{R}_+ \times [0, t]$, by a 3-dimensional grid $\{t_i\}_{i=1}^{N_t} \times \{y_j\}_{j=1}^{N_y} \times \{s_k\}_{k=1}^{N_s}$ with mesh sizes $(\Delta t, \Delta y, \Delta t)$, where $N_t, N_y, N_s > 0$ are the partition sizes for each dimension.⁷ For ease of notation, write $v_{i,j} = v(t_i, y_j)$ and $h_{i,j,k} = h(t_i, y_j, s_k)$, with the relationship $v_{i,j} = h_{i,j,i}$ for all $i = 1, \dots, N_t$ and $j = 1, \dots, N_y$. The boundary conditions are given by

$$v_{N_t,k} = g(e^{y_{N_t}}) \quad (5.59)$$

$$h_{N_t,j,k} = \beta(t_{N_t} - s_k)g(e^{y_j}) \quad (5.60)$$

$$h_{i,0,k} = \beta(t_i - s_k)g(e^{y_0}) \quad (5.61)$$

⁷Note that the state dimension $\mathbb{R}_+$ is unbounded. We consider a trimmed state space $[0, X_{cap}]$ for some large value $X_{cap} > 0$ such that it is very unlikely that X will breach X_{cap} in a finite horizon $[0, T]$. The partition $\{y_j\}_{j=1}^{N_y}$ is performed on the trimmed state space. Also, for each $t \in [0, T]$, the domain of s is $[0, t]$. But in practice, we partition t and s on the same domain $[0, T]$ and consider an equal mesh size Δt for t and s (so $N_t = N_s$) for technical convenience. Then, we fix $h(t_i, y_j, s_k)$ to zero whenever $s_k > t_i$, for all $i = 1, \dots, N_t, j = 1, \dots, N_y$ and $k = 1, \dots, N_s$.

$$h_{i,N_y,k} = \beta(t_i - s_k)g(e^{y_{N_y}}), \quad (5.62)$$

for $i = 0, \dots, N_t, j = 0, \dots, N_y$ and $k = 0, \dots, N_s$. On the grid, the infinitesimal generators $\mathcal{L}f$ and $\mathcal{L}V$ are respectively approximated by

$$\begin{aligned} \mathcal{L}h_{i,j,k} &= \frac{h_{i+1,j,k} - h_{i,j,k}}{\Delta t} + \left(\mu - \frac{1}{2}\sigma^2\right) \frac{h_{i,j+1,k} - h_{i,j-1,k}}{2\Delta y} + \frac{1}{2}\sigma^2 \frac{h_{i,j+1,k} - 2h_{i,j,k} + h_{i,j-1,k}}{(\Delta y)^2} \\ &= \frac{h_{i+1,j,k} - h_{i,j,k}}{\Delta t} + \left(\frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) h_{i,j+1,k} \\ &\quad - \frac{\sigma^2}{(\Delta y)^2} h_{i,j,k} + \left(\frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) h_{i,j-1,k} \end{aligned} \quad (5.63)$$

$$\begin{aligned} \mathcal{L}v_{i,j} &= \frac{v_{i+1,j} - v_{i,j}}{\Delta t} + \left(\frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) v_{i,j+1} \\ &\quad - \frac{\sigma^2}{(\Delta y)^2} v_{i,j} + \left(\frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) v_{i,j-1}. \end{aligned} \quad (5.64)$$

Consequently, for $i = 0, \dots, N_t, j = 0, \dots, N_y$ and $k = 0, \dots, N_s$, the PDE of f (5.31) is approximated by

$$\begin{aligned} &\frac{h_{i+1,j,k} - h_{i,j,k}}{\Delta t} + \left(\frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) h_{i,j+1,k} \\ &\quad - \frac{\sigma^2}{(\Delta y)^2} h_{i,j,k} + \left(\frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y}\right) h_{i,j-1,k} \\ &\quad + K(\beta(t_i - s_k)g(e^{y_j}) - h_{i,j,k})\mathbb{I}(\beta(t_i - s_k)g(e^{y_j}) > h_{i,j,k}) = 0, \end{aligned} \quad (5.65)$$

where $\mathbb{I}(\cdot)$ is the indicator operator that equals zero if the condition is met. After rearranging and merging common terms, this simplifies to

$$a_j^h h_{i,j-1,k} + b_j^h h_{i,j,k} + c_j^h h_{i,j+1,k} = d_j^h, \quad (5.66)$$

where

$$a_j^h = \frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \quad (5.67)$$

$$b_j^h = -\frac{\sigma^2}{(\Delta y)^2} - \frac{1}{\Delta t} - K\mathbb{I}(\beta(t_i - s_k)g(e^{y_j}) > h_{i,j,k}) \quad (5.68)$$

$$c_j^h = \frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \quad (5.69)$$

$$d_j^h = -\frac{h_{i+1,j,k}}{\Delta t} - K\beta(t_i, s_k)g(e^{y_j})\mathbb{I}(\beta(t_i - s_k)g(e^{y_j}) > h_{i,j,k}). \quad (5.70)$$

Likewise, for $i = 0, \dots, N_t, j = 0, \dots, N_y$ and $k = 0, \dots, N_s$, the PDE of V (5.30) is approximated by

$$\begin{aligned} & \frac{v_{i+1,j} - v_{i,j}}{\Delta t} + \left(\frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \right) v_{i,j+1} \\ & - \frac{\sigma^2}{(\Delta y)^2} v_{i,j} + \left(\frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \right) v_{i,j-1} \\ & + K(g(e^{y_j}) - v_{i,j})\mathbb{I}(g(e^{y_j}) > v_{i,j}) - \frac{h_{i,j,i} - h_{i,j,i-1}}{\Delta s} = 0, \end{aligned} \quad (5.71)$$

which simplifies to

$$a_j^v v_{i,j-1} + b_j^v v_{i,j} + c_j^v v_{i,j+1} = d_j^v, \quad (5.72)$$

where

$$a_j^v = \frac{\sigma^2}{2(\Delta y)^2} - \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \quad (5.73)$$

$$b_j^v = -\frac{\sigma^2}{(\Delta y)^2} - \frac{1}{\Delta t} - K\mathbb{I}(g(e^{y_j}) > v_{i,j}) \quad (5.74)$$

$$c_j^v = \frac{\sigma^2}{2(\Delta y)^2} + \frac{\mu - \frac{1}{2}\sigma^2}{2\Delta y} \quad (5.75)$$

$$d_j^v = -\frac{v_{i+1,j}}{\Delta t} - Kg(e^{y_j})\mathbb{I}(g(e^{y_j}) > h_{i,j,i}) + \frac{h_{i,j,i} - h_{i,j,i-1}}{\Delta t}. \quad (5.76)$$

Note that $a_j^h = a_j^v$ and $c_j^h = c_j^v$, and they are constant. Also, b_j^h and b_j^v are equal except for the indicator $\mathbb{I}$. Using the tridiagonal matrix algorithm (Press et al., 2007), we solve for $v_{i,j}$ and $h_{i,j,k}$ on the 3-dimensional grid, i.e., the discretized approximation of V and f . The numerical algorithm involves repeatedly updating $v_{i,j}$ and $h_{i,j,k}$ (for all i, j, k) in an alternating way, which is summarized below in Algorithm 5.

Algorithm 5: Solving for V and f via finite difference method

Initialize mesh sizes $(\Delta t, \Delta y, \Delta s)$, partition sizes (N_t, N_y, N_s) , number of iterations N_{iter} , early stopping threshold $\epsilon > 0$, initial values of $v_{i,j,k}^{(0)}$ and $h_{i,j,k}^{(0)}$ with boundary conditions satisfying (5.59) to (5.62).

```
for  $l = 0, \dots, N_{iter} - 1$  do
  for  $i = N_t - 1, \dots, 0$  do
    for  $k = 0, \dots, i - 1$  do
      Compute (5.66) for each  $j = 1, \dots, N_y$ , with  $h_{i,j,k}$  replaced by  $h_{i,j,k}^{(l)}$  in
      the coefficients.
      Solve for  $h_{i,j,k}^{(l+1)}$  via the tridiagonal matrix algorithm.
    Compute (5.72) for each  $j = 1, \dots, N_y$ , with  $v_{i,j}$  replaced by  $v_{i,j}^{(l)}$  and  $h_{i,j,k}$ 
    replaced by the just-solved  $h_{i,j,k}^{(l+1)}$  in the coefficients.
    Solve for  $v_{i,j}^{(l+1)}$  via the tridiagonal matrix algorithm.
  Check convergence and stop early if  $\max_{i,j} |v_{i,j}^{(l+1)} - v_{i,j}^{(l)}| < \epsilon$ .
```

5.5 Numerical Results

In this section, we consider an optimal exercise problem of an American put option with discounting, on a finite horizon ($T = 1$). We apply our RL algorithm to learn the optimal exploratory exercise strategies, under exponential and non-exponential discounting settings. We use the finite difference algorithm to approximate the “grounding true” value functions, which helps us to understand the quality of RL training. Beyond that, we compare the exercise strategies under exponential and non-exponential discounting settings, which empirically verifies our findings from Proposition 5.3.1 and Proposition 5.3.2, that non-exponential discounting investors are more patient and thus tend to exercise the options at later times for a potentially better price.

We define an American put option problem as an optimal stopping problem with discounting, namely (5.22), which we repeat here

$$\max_{\tau \in \mathcal{T}_{[0,T]}} \mathbb{E} [\beta_{\tau} (S_{\text{strike}} - X_{\tau})^+], \quad (5.77)$$

where S_{strike} is the strike price. X_t is the stock price at time t driven by

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \text{ with } X_0 = x_0, \quad (5.78)$$

where $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ are respectively the growth rate and volatility of the stock. The terminal payoff is replaced by the payoff of a put option, i.e., $g(x) := (S_{\text{strike}} - x)^+$. For this simulation study, we set $(\mu, \sigma, S_{\text{strike}}, x_0) = (0.06, 0.4, 40, 40)$. Note that the true values of (μ, σ) are not accessible to the RL model, whereas the finite difference algorithm relies on the knowledge of the true values of (μ, σ) . Finally, we set the mesh size $\Delta t = \frac{1}{50}$ for the planning horizon $[0, 1]$, hence an investor has $\frac{T}{\Delta t} = 50$ time steps to decide whether to exercise the put option.

We use gamma discounting as an example of non-exponential discounting. Adopting the definition of gamma discounting in (Ebert et al., 2020),

$$\beta_t^{\text{gamma}} := (1 + \alpha t)^{-\frac{\beta}{\alpha}}, \quad (5.79)$$

with $\alpha, \beta > 0$ and for $t \in [0, T]$. We note that β_t^{gamma} can also be expressed as an expectation of exponential discounting with random discount rates following a Gamma distribution F , namely,

$$\beta_t^{\text{gamma}} = \int_0^\infty e^{-\rho t} dF(\rho). \quad (5.80)$$

Denote by μ_F, σ_F the mean and standard deviation of this Gamma distribution F . According to (Weitzman, 2001), the relations between (α, β) and (μ_F, σ_F) are

$$\alpha = \frac{\sigma_F^2}{\mu_F} = 0.0225 \quad (5.81)$$

$$\beta = \mu_F = 0.04, \quad (5.82)$$

where $\mu_F = 0.04$ and $\sigma_F = 0.03$ are selected based on a survey over 2000 economists.

For a fair comparison to gamma discounting, we consider exponential discounting $\beta_t^{\text{exp}} = e^{-\rho t}$ with the discount rate equals to the mean of the Gamma distribution, i.e., $\rho = \mu_F = 0.04$. According to Proposition 2 and Theorem 2 of (Ebert et al., 2020), we have $\beta_t^{\text{gamma}} \geq \beta_t^{\text{exp}}$, for all $t \in [0, T]$. Beyond that, according to Proposition 5.3.1, we expect the stopping region for the chosen gamma discounting to be smaller than the stopping region for the chosen exponential discounting. Figure 5.2 confirms this empirically. On the left, the gamma discounting function dominates the exponential discounting function on $(0, 1]$, as they both start from 1 at $t = 0$. On the right, the decision boundary for exponential discounting is numerically solved via the finite difference method addressed in (Dai et al., 2024), whereas the decision boundary of gamma discounting is calculated based on the approximated value functions via our finite difference algorithm (Algorithm 5). Clearly, the decision boundary of the gamma discounting lies below that of the exponential discounting, implying a smaller stopping region under gamma (non-exponential) discounting.

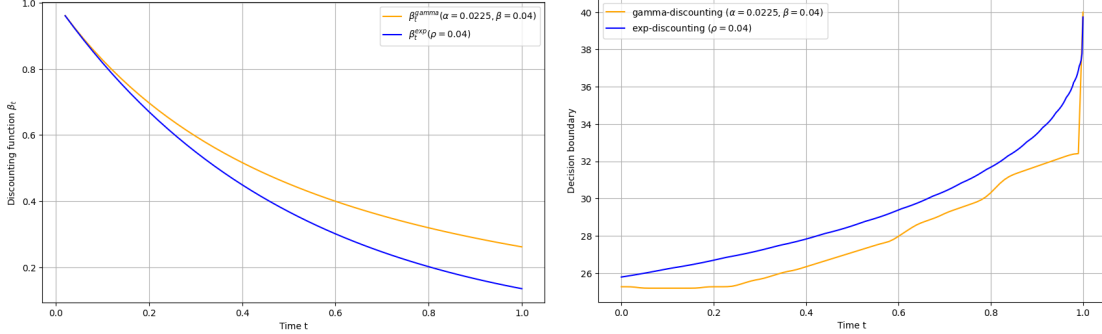


Figure 5.2: Discounting functions (left) and their decision boundaries (right).

Next, we configure a neural network (NN) to learn the value function. Following (Dai et al., 2024), we configure a separate NN to learn the value function at each discretized time point $l = 1, \dots, 50$. Denote by V_{θ_l} the NN-approximated value function at time t_l , for $l = 1, \dots, 50$. Each V_{θ_l} takes the stock price X_{t_l} and time t_l as input and returns an approximation of $V(t_l, X_{t_l})$. Note that $\{V_{\theta_l}\}_{l=1}^{N_t}$ are not connected. We summarize the NN configuration in Table 5.1.

Layer type	Dimension	Is shared?
MLP	(1, 16)	No
MLP	(16, 32)	No
MLP	(32, 64)	No
MLP	(64, 64)	No
MLP	(64, 32)	No
MLP	(32, 16)	No
MLP	(16, 1)	No

Table 5.1: RL model configuration. Total number of trainable parameters: 481,171.

Figure 5.3 illustrates the decision boundaries learned via the RL algorithm (Algorithm 4) and the finite difference algorithm (Algorithm 5), as described in Section 5.4. Since the numerical algorithm employs the true values of the market parameters (μ, σ) , we use the induced decision boundary (called the free boundary) as the benchmark, shown as the blue curves in Figure 5.3. To find the decision boundary of the RL algorithm, we simulate 100 stock price trajectories $\{X_{t_l}^{(n)}\}_{l=1, \dots, 50}^{n=1, \dots, 100}$ and compute the stopping policies $\pi_{t_l}^{(n)}$ at each time step l on each stock price trajectory n . We mark a point $X_{t_l}^{(n)}$ as “exercise” if $\pi_{t_l}^{(n)} \geq 50\%$.

and mark “do not exercise” otherwise. The scatters in Figure 5.3 demonstrate the 100 simulated stock price trajectories, with “exercise” in green and “do not exercise” in red. The edge of the red and the green scatters is approximately the decision boundary learned by the RL algorithm, called the *empirical free boundary*. Clearly, the *free boundaries* closely match the *empirical free boundaries* under both exponential discounting and non-exponential discounting settings.

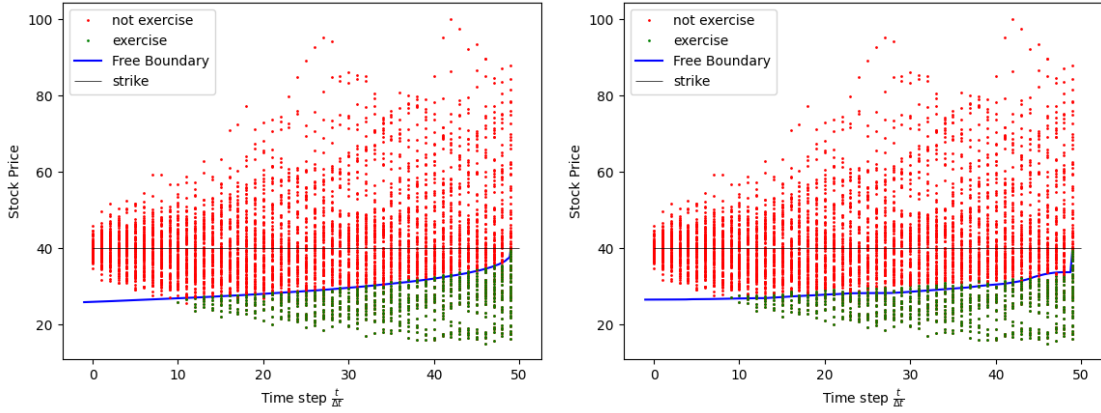


Figure 5.3: The decision boundaries learned via the RL algorithm are characterized by the edges between the “exercise” (green) and “do not exercise” (red) points, under exponential discounting (left) and non-exponential discounting (right) settings. The free boundaries (blue curves) are numerically solved via the finite difference method and are considered as the benchmark.

We also compare the algorithms and the effect of discounting functions in terms of option payoffs. For each discounting function, we determine early exercise on a given stock price trajectory based on each algorithm and calculate the discounted option payoff. We term the payoff that is realized based on the RL algorithm as an *empirical payoff*, and the payoff that is realized based on the finite difference algorithm as a *theoretical payoff*. Since the *theoretical payoff* is considered as the benchmark of the *empirical payoff* for a given stock price trajectory, we calculate the *relative error* as $\left| \frac{\text{empirical} - \text{theoretical}}{\text{theoretical}} \right| \times 100\%$.

Table 5.2 reports the mean *empirical payoff* and mean *theoretical payoff* calculated from 100 independent batches of stock prices, each batch containing 100 independently simulated stock price trajectories, for a total of 10,000 stock price trajectories. The payoffs are reported in present values at time $t = 0$, discounted by the corresponding exponential and gamma discounting functions. Note that the relative errors under both exponential

and non-exponential discounting settings are small (near 0.15%). This indicates that the RL algorithm learns the optimal exercise policy fairly well, despite having no access to the underlying stock price dynamics. Furthermore, the mean payoffs of non-exponential discounting, both empirical and theoretical, are greater than those of exponential discounting.

discounting type	exponential discounting ($\rho = 0.04$)	non-exponential discounting ($\alpha = 0.0225, \beta = 0.04$)
mean empirical payoff	5.3578	5.4829
mean theoretical payoff	5.3660	5.4911
relative error	0.1539%	0.1495%
average exercise time	0.8029	0.9195

Table 5.2: Discounted option payoffs.

Beyond that, for each stock price trajectory, we also calculate the exit time following the optimal exercise policies under exponential and non-exponential discounting. Table 5.2 shows that a non-exponential discounting investor tends to exercise the option at a later time on average, compared to an exponential discounting investor. We draw the empirical densities and CDFs of the exit times for both exponential and non-exponential discounting in Figure 5.4. Under both discounting functions, the majority exit time is at the maturity of the option ($t = T = 1$). In spite of that, the empirical density of exit times for exponential discounting investors puts more mass on earlier times ($t < T = 1$) than that of non-exponential discounting investors. Likewise, the empirical CDF of the exit times under non-exponential discounting is also to the right of the empirical CDF under exponential discounting.

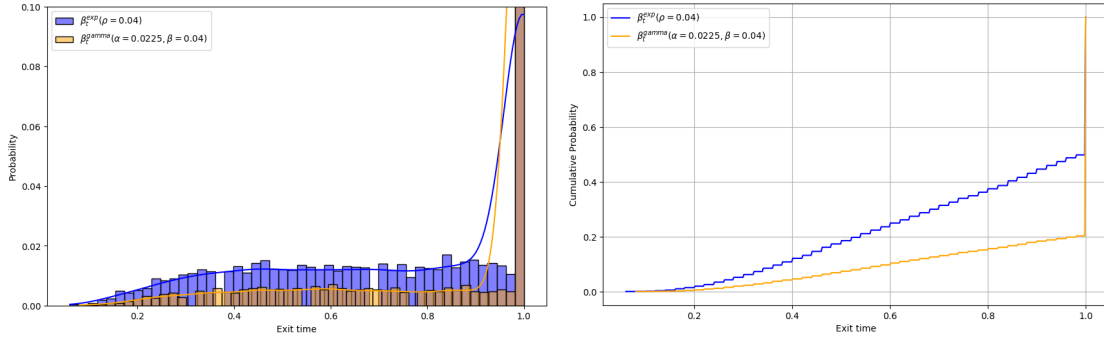


Figure 5.4: Empirical densities (left) and CDFs (right) of exit times for exponential and non-exponential discounting.

Finally, we compute the average exit time of the 100 stock price trajectories for each of the 100 batches, resulting in 100 average exit times under each discounting function. The empirical densities of the average exit times are illustrated in Figure 5.5, which reiterates the fact that non-exponential discounting leads to a later stopping time on average. Our simulation results agree with our findings in Proposition 5.3.1 and Proposition 5.3.2.

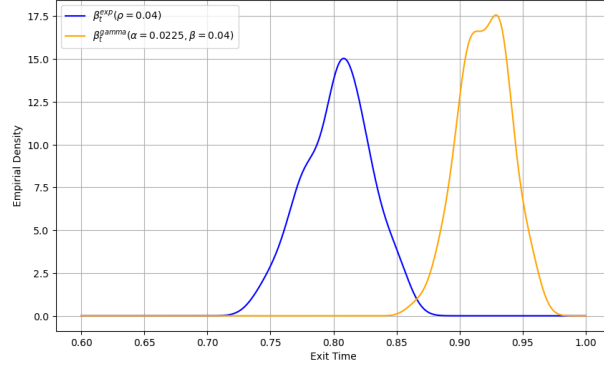


Figure 5.5: Empirical densities of average exit times for exponential and non-exponential discounting.

5.6 Conclusion

This chapter generalizes (Dai et al., 2024) by accounting for a general, potentially non-exponential, discounting function multiplying the payoff function of the optimal stopping problem. We apply the penalty approximation technique as described in (Dai et al., 2024) to the free-boundary PDEs of the original discounted optimal stopping problem (Problem (5.22)), transforming them to a system of penalized PDEs, that can represent the HJB of a standard stochastic control problem. Then, following the exploratory extension argument from (H. Wang et al., 2020), we further extend the discounted optimal stopping problem to the discounted exploratory stopping problem (Problem (5.24)), as well as the corresponding EHJB equations, (5.26) and (5.27). We solve for a semi-analytical Bernoulli stopping policy (5.32) that depends on the value function. While our stopping policy takes the same form as the optimal stopping policy derived in (Dai et al., 2024), we remind the readers that the two stopping policies are different by design and by the type of optimality. Dai et al. (2024) solved a time-consistent control problem, yielding a precommitted stopping policy that is globally optimal at the initial time, whereas this chapter studies a time-inconsistent control problem due to the presence of the non-exponential discounting. Hence, our stopping policy

is locally optimal in a game-theoretical equilibrium sense, so that any deviation from this policy will make the agent worse off.

Furthermore, we examine the effect of discounting on the optimal stopping problem. We suppose the non-exponential discounting function β^{nonexp} can be expressed as a weighted sum of exponential discounting functions with random discounting rates following some weighting distribution. To establish a fair comparison, we further set the discounting factor of the exponential discounting function β^{exp} as the mean of the weighting distribution. Then, we solve the stopping problems under the classical (non-exploratory) and the exploratory formulations, with exponential and non-exponential discounting. Consequently, we observe two interesting results. First, under the classical formulation, the stopping region with non-exponential discounting is smaller than that with exponential discounting. Second, under the exploratory formulation, the optimal stopping policy (defined as the probability of stopping) with non-exponential discounting is lower than that with exponential discounting. Intuitively, our result depicts an agent with higher patience (and potentially more ambition) modeled by non-exponential discounting, willing to wait for a deferred but possibly better payoff. In contrast, an agent with exponential discounting has a higher cost of waiting, hence being less patient and tending to secure an earlier payoff.

Based on the theoretical results, we develop an RL algorithm to learn the value function. In the RL algorithm, we parameterize the value function with a neural network, as the value function of a finite horizon stopping problem is very challenging to solve analytically. Besides, this also yields the Bernoulli stopping policy according to its semi-analytical form (5.32). We follow the policy evaluation method from (Jia & Zhou, 2022a) and define a stochastic process based on the EHJB equations, that is proven to be a martingale when the stopping policy is an equilibrium policy, or equivalently, when the neural network correctly approximates the value function. Thereby, we derive a martingale loss function that leverages the martingale property from the so-defined process, which effectively trains the neural network to learn the value function.

On the other hand, we also develop a finite difference algorithm. Due to the time-inconsistency, we need to solve for both the value function and the auxiliary value function. Note from their probabilistic interpretations that the auxiliary value function is effectively the value function discounted to an earlier time point. So, while the value function is a function of the time-state pair, the auxiliary value function is a function of the time-state-time triplet. On a 3-dimensional grid, we use the finite difference method and iteratively update the value function and the auxiliary value function in turn.

As summarized in Algorithms 4 and 5, the RL algorithm does not require the full knowledge of the underlying model dynamics, namely the coefficients of the state process

(5.2), which are required by the finite difference algorithm. Therefore, we are careful with our language when describing the roles of the two algorithms — the RL algorithm “learns” the value function via interactions with the underlying dynamics, while the finite difference algorithm “solves” for the value function given the underlying dynamics.

Finally, we verify our findings through a simulation study. We show empirically that the free boundary under gamma discounting, a common non-exponential discounting, lies below the free boundary under a comparable exponential discounting. Hence, the stopping region (the area below the free boundary curve) of gamma discounting is smaller than that of exponential discounting. Moreover, in an American put option example, a gamma discounting investor on average exercises the option at a later time than an otherwise equivalent exponential discounting investor, which indicates the lower chance of exercising the option and higher level of patience of a gamma discounting investor. Beyond that, our simulation study also demonstrates the effectiveness of the RL algorithm. A trained RL investor is able to follow the free boundary derived from the finite difference algorithm, by repeatedly deciding when to exercise an American put option over batches of simulated stock price trajectories. The mean realized payoff of the trained RL investor is close to the mean theoretical payoff calculated through the finite difference algorithm, with low relative error around 0.15%. Such performance is consistent under exponential and gamma discounting. Considering the finite difference algorithm as the benchmark, the RL algorithm effectively and accurately learns the value function, as well as the stopping policy, without access to the underlying model dynamics.

Chapter 6

Conclusion and Future Works

This thesis applies continuous time RL techniques to solve four stochastic control problems in the exploratory control framework, which can be broadly categorized into three well-studied finance problems — the mean-variance portfolio optimization problem (Chapters 2 and 3), the optimal investment-consumption problem (Chapter 4) and the optimal stopping problem with application in exercising American-type options (Chapter 5). The main contributions of this thesis to the literature can be seen from three aspects.

First and foremost, we follow the recent trend from the stochastic control literature that applies RL techniques to advance classical solutions of standard control problems, and fill the gaps between the control literature and the RL literature at four critical points, i.e., mean-variance with Markovian regime-switching dynamics, mean-variance with jump-diffusion controlled dynamics, optimal investment-consumption and optimal stopping. In each problem, we extend the classical control setup to an exploratory control framework (following (H. Wang et al., 2020)), randomizing the control from a deterministic function of time and state to a probability distribution over the action space conditioned on a time-state pair. We solve, either analytically or semi-analytically, for the optimal policy and the corresponding value function, from the derived HJB equations. We also provide verification theorems and policy improvement (or policy iteration) theorems, as well as the subsequent lemmas and propositions, that complement the theoretical foundations of the exploratory extensions to each classical control problem we study.

Second, we develop RL algorithms to learn the optimal policies and the corresponding value functions, without requiring a full knowledge on the model dynamics of the environment. Our RL algorithms are *value-based*, meaning they learn the optimal policy through estimating the value function, in contrast to *policy-based* algorithms, which directly op-

timize the policy by back-propagating gradient through it; see (Hambly et al., 2023) for categorization of *value-based* and *policy-based* RL algorithms and more examples. Value-based RL algorithms fit exploratory control framework better, as the optimal policies are deduced from the value function based on the HJB equations, regardless of whether the optimal policies can be solved in explicit forms. We also propose effective loss functions following arguments in (Jia & Zhou, 2022a, 2023), which extensively leverage the martingale properties from the theoretical solutions¹. As discussed in (Jia & Zhou, 2022a) and also empirically shown in Chapter 2, our loss functions are more suitable for the exploratory control problems that we study, compared to temporal-difference learning, a widely applied method in the RL literature (Sutton & Barto, 2018). Moreover, under proper configurations in Chapters 2, 3 and 4, our RL models not only learn the optimal policies and the value functions accurately, but also calibrates the true diffusion model parameters of the dynamic environment, without any parameter estimation involved.

Third, we overcome problem-specific difficulties with transferable techniques from the literature, which can be recalled when dealing with similar situations. In Chapter 2, we aim to solve for a precommitted strategy, hence considering the Lagrangian dual of the original mean-variance problem, as also studied in (H. Wang & Zhou, 2020). As a result, this problem is solved via the standard Dynamic Programming Principle (DPP) under a time-consistent setting. In Chapter 3, we directly deal with the non-linearity² of the expectation in the mean-variance objective function and the time-inconsistency associated to it, where the standard DPP fails. We adopt an equilibrium approach from (Björk et al., 2021) and solve for an investment policy that is locally optimal in a game-theoretical equilibrium sense. In Chapter 4, we tackle the dual-control problem by considering an equivalent auxiliary optimization problem, from which the solved equilibrium policy can be transformed from a joint distribution over the investment and consumption space to a product of two marginal distributions of investment and consumption. Such a separation of the joint-policy not only simplifies the RL algorithm design but also demonstrates the practical intuition of the relationship between investment and consumption — an investor’s consumption is a more inelastic demand that should be independent of her investment decisions, despite more intrinsically they are both constrained by the investor’s total wealth. In Chapter 5, we address the free-boundary PDEs of the optimal stopping problem via a penalty approximation method similar to (Dai et al., 2024; Peng et al., 2024), which re-

¹When the market is extremely general and complex such that a true martingale is not easily derivable, we have to instead consider more direct methods such as neural network approximation of the value function or policy gradient algorithms (Sutton et al., 1999), although this is beyond the scope of this thesis.

²The non-linearity arises from the variance term in the mean-variance objective, as variance can be expressed as a quadratic function of expectation.

formulates them into a system of penalized PDEs, that can be considered as the extended HJB equations of a standard time-inconsistent control problem. Such reformulation enables us to apply the exploratory extension and the already established RL methods for exploratory control problems.

As a high level summary, we remark that the major works of this thesis can be viewed and discussed in four additional perspectives — time consistency, preference type, control space and RL model parameterization — besides categorizing them into three finance applications.

From the time consistency perspective, Chapter 2 studies a time-consistent control problem, whereas Chapters 3, 4 and 5 study time-inconsistent control problems. As a result, Chapter 2 solves for a unique precommitted strategy that is globally optimal at the initial time. However, with the presence of time-inconsistency, an agent’s preference may vary over time. Consequently, a policy that is optimal right now can become suboptimal when evaluated at a later time. Thus, we seek for a policy that an agent is incentivized to stick with throughout the planning horizon, as any small deviation from this policy will make the agent worse-off. Such a policy is known as an equilibrium policy, as well-documented by (Björk et al., 2021), imagining an agent competes with her future selves and winds up with a game-theoretical equilibrium. Yet, as discussed by (Dai et al., 2023), the uniqueness of an equilibrium solution to a time-inconsistent control problem remains as an open question. Hence, the policies derived in Chapters 3, 4 and 5 are locally optimal in a game-theoretical equilibrium sense, but are not necessarily the only solution to their problems.

From the perspective of preferences, we note from economics work in utility theory such as (Myers, 1968) that intertemporal preferences can be decomposed into *state preferences*, reflecting an agent’s attitudes toward her temporal state such as wealth level, and *time preferences*, capturing an agent’s satisfactions toward a reward to be received at different times (such as one dollar today versus tomorrow). Correspondingly, Björk et al. (2021) described two possible sources of time-inconsistency from the objective function of a stochastic control problem, a non-linear term of expectation and a non-exponential discounting, which can be interpreted as modeling the state and time preferences, respectively. Among the time-inconsistent control Chapters (3, 4 and 5), Chapter 3 addresses the state preferences as it tackles the non-linearity in expectation (the variance term) in the mean-variance objective function, whereas Chapters 4 and 5 focus on time preferences due to the incorporation of a non-exponential discounting function.

From the control space perspective, Chapters 2 and 3 study a single control (investment) defined on $\mathbb{R}$, Chapter 4 addresses a dual-control (investment and consumption) defined

on $\mathbb{R} \times \mathbb{R}_+$, and Chapter 5 works on a single control defined on a finite set $\{\text{stop}, \text{continue}\}$. Under the exploratory control framework, the optimal policy follows the Boltzmann family of distributions, as a result of incorporating entropy regularization. As the Hamiltonian is quadratic in the investment control, the optimal investment policies in Chapters 2, 3 and 4 are Gaussian distributions. As the Hamiltonian is logarithmic in the consumption control, the optimal consumption policy in Chapter 4 becomes a Gamma distribution. As the control space is restricted to exactly two options, the optimal stopping policy in Chapter 5 yields a Bernoulli distribution.

From the RL model parameterization perspective, all of the policies in Chapters 2, 3, 4 and 5 share the same set of parameters with their corresponding value functions, which is different from a general Actor-Critic algorithm as described in (Sutton & Barto, 2018; Jia & Zhou, 2022a) that parametrizes the policy and the value function independently. This shared-parameters configuration is a direct implication of our theoretical results under the exploratory control framework, where the policies depend on the corresponding value functions according to the HJB equations. This design significantly reduces the number of trainable parameters and simplifies the RL model training process, from training the policies and value functions separately (and often alternately) to training the value functions alone and updating the policies. Furthermore, the RL models in Chapters 2, 3 and 4 are parameterized by the model parameters of the environment, such as the mean and volatility parameters of the stock price dynamics in Chapter 2. Because the value functions can be analytically or semi-analytically solved, hence so are the policies. As a result, the RL models in Chapters 2, 3 and 4 not only learn the optimal policies and the corresponding value functions, but also accurately calibrate the “grounding true” values of the dynamic environment’s model parameters, without any estimation involved. Nevertheless, the RL model in Chapter 5 is parameterized by a neural network, as the value function of the optimal stopping problem considered in this thesis cannot be solved analytically in finite horizon. Broadly, the RL modeling pipeline and training algorithms in this thesis are implementable with and without analytical solutions. In Chapters 2, 3 and 4, suppose the market dynamics are generalized to more complicated stochastic processes or higher dimensions (in a multi-asset setup) where closed form solutions of the value functions are no longer solvable. One can consider to parameterize the value functions with neural networks instead of with the market model’s parameters. While neural network-based RL models’ parameters lose practical meanings, the RL algorithms are still functional and are able to achieve meaningful performance on practical training tasks as illustrated in Chapter 5.

Finally, with some humble thoughts, we devote the remainder of this thesis to three prospects of future extensions to our works.

Separation of market stochasticity and policy exploration. In Chapter 2, we mention three major sources of randomness in the EMVRS formulation — the market dynamics, the exploration of the investment strategy and the market regimes. While the market regime is modeled on a Markov Chain, the market dynamics (2.1) and the portfolio value process (2.3) are driven by the same Brownian motion. Note that we have a similar formulation in Chapters 3, 4 and 5 as well. This could potentially be generalized to different Brownian motions with some covariance, separating the market stochasticity from exploration randomness; see (Dai et al., 2023) for the stochasticity separation with different Brownian motions.

Bayesian approach to learn the model parameters. We acknowledge the possibility of learning the model parameters of the dynamic environment via a Bayesian approach, in addition to RL. Because our RL model parameters in Chapters 2, 3 and 4 all have practical meanings and the number of parameters is small. One potential way is to apply a filtering method (Fulop, 2011) to iteratively learn the posterior distribution of each model parameter. The stochasticity of the exploratory control thereby has two sources, one from the distribution-valued policy and the other from the posterior distribution of the model parameters. Besides, one can also employ Bayesian RL techniques to learn the posterior distributions of the model parameters, instead of learning a single estimate as in the classical RL approaches; see (Vlassis et al., 2012; Ghavamzadeh et al., 2015) for examples of Bayesian RL methods.

Learning the reward function. In Chapters 4 and 5, we assume that the discounting function is given and we implement the RL algorithm on a simulation study using gamma discounting with given parameters; see for example Section 4.6. One possible extension is to set parameters of the discounting function free, and learn the true discounting parameters with the market parameters. This frames into a new type of control problem with uncertain objective function, which can potentially be tackled via Inverse Reinforcement Learning (IRL) (Ng & Russell, 2000). IRL is a framework within machine learning aimed at inferring the underlying reward function that governs an agent’s observed behavior. For instance, one can apply IRL methods to derive the implicit reward structure behind successful investors based on their historical investment activities, in order to replicate or improve trading algorithms. Unlike traditional RL, which optimizes a policy given a pre-defined reward structure, IRL seeks to uncover the hidden motivations behind the actions of an agent, enabling insights into decision-making processes or allowing replication of observed behavior in similar environments. When the discounting parameters are masked, the reward function of an investment problem is partially hidden. Yet, application of IRL

in financial problems are very limited in the literature, in spite of its potential in practice. See (Ng & Russell, 2000) for the original introduction of IRL and (Adams et al., 2022; Deshpande et al., 2025) for more comprehensive review of recent advances and applications of IRL.

We conclude here with a remark that the above three directions are not exclusive of each other. For instance, one can always adopt the idea of separating the stochasticity of the dynamic environment and the policy exploration with different Brownian motions, in a future exploration of learning the posterior distributions of model parameters via Bayesian RL. As the author of this thesis, I thank the readers for reading to this final page and sincerely wish that it has clearly addressed the research questions and may serve as a modest inspiration for future scholars.

Bibliography

- Achdou, Y., & Pironneau, O. (2005). *Computational methods for option pricing*. SIAM.
- Adams, S., Cody, T., & Beling, P. A. (2022). A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55(6), 4307–4346.
- Amin, K. I. (1993). Jump diffusion option valuation in discrete time. *The Journal of Finance*, 48(5), 1833–1863.
- Andersen, L., & Andreasen, J. (2000). Jump-diffusion processes: Volatility smile fitting and numerical methods for option pricing. *Review of Derivatives Research*, 4, 231–262.
- Ang, A., & Bekaert, G. (2004). How regimes affect asset allocation. *Financial Analysts Journal*, 60(2), 86–99.
- Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4), 1137–1187.
- Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6), 26–38.
- Bae, G. I., Kim, W. C., & Mulvey, J. M. (2014). Dynamic asset allocation for varied financial markets under regime switching framework. *European Journal of Operational Research*, 234(2), 450–458.
- Barro, R. J. (1999). Ramsey meets Laibson in the neoclassical growth model. *Quarterly Journal of Economics*, 114(4), 1125–1152.
- Bates, D. S. (1991). The crash of '87: Was it expected? The evidence from options markets. *The Journal of Finance*, 46(3), 1009–1044.
- Bayraktar, E., & Egami, M. (2008). Optimizing venture capital investments in a jump diffusion model. *Mathematical Methods of Operations Research*, 67(1), 21–42.
- Becker, S., Cheridito, P., & Jentzen, A. (2020). Pricing and hedging american-style options with deep learning. *Journal of Risk and Financial Management*, 13(7), 158.
- Bender, C., & Thuan, N. T. (2026). Continuous time reinforcement learning: A random measure approach. *Stochastic Processes and their Applications*, 194, 104848. <https://doi.org/https://doi.org/10.1016/j.spa.2025.104848>

- Bender, C., & Thuan, N. T. (2023). Entropy-regularized mean-variance portfolio optimization with jumps. *arXiv preprint arXiv:2312.13409*.
- Björk, T., Khapko, M., & Murgoci, A. (2016). A theory of markovian time inconsistent stochastic control in continuous time. *Available at SSRN 2727480*.
- Björk, T., Khapko, M., & Murgoci, A. (2021). *Time-inconsistent control theory with finance applications*. Springer.
- Björk, T., & Murgoci, A. (2010). A general theory of markovian time inconsistent stochastic control problems. *Available at SSRN 1694759*.
- Björk, T., & Murgoci, A. (2014a). On time-inconsistent stochastic control in continuous time. *Finance and Stochastics*, 18(3), 545–592.
- Björk, T., & Murgoci, A. (2014b). A theory of markovian time inconsistent stochastic control in discrete time. *Finance and Stochastics*, 18, 545–592.
- Brennan, M. J. (1998). The role of learning in dynamic portfolio decisions. *The Review of Financial Studies*, 11(3), 525–548.
- Bruss, F. T. (2000). Sum the odds to one and stop. *Annals of Probability*, 1384–1391.
- Carr, P., & Mayo, A. (2007). On the numerical evaluation of option prices in jump diffusion processes. *European Journal of Finance*, 13(4), 353–372.
- Chen, Y. M., Li, B., & Saunders, D. (2025a). Exploratory investment-consumption problem with non-exponential discounting. *SIAM Journal on Financial Mathematics (Under Review)*.
- Chen, Y. M., Li, B., & Saunders, D. (2025b). Exploratory mean-variance portfolio optimization with regime-switching market dynamics. *International Journal of Theoretical and Applied Finance*, 28(05n06), 2550014. <https://doi.org/10.1142/S0219024925500141>
- Chen, Y. M., Li, B., & Saunders, D. (2025c). Exploratory mean-variance portfolio optimization: An equilibrium approach. *arXiv preprint arXiv:2512.09224*.
- Chiu, M. C., & Li, D. (2006). Asset and liability management under a continuous-time mean-variance optimization framework. *Insurance: Mathematics and Economics*, 39(3), 330–355.
- Chow, Y.-S., Moriguti, S., Robbins, H., & Samuels, S. M. (1964). Optimal selection based on relative rank (the “secretary problem”). *Israel Journal of Mathematics*, 2(2), 81–90.
- Clift, S. S., & Forsyth, P. A. (2008). Numerical solution of two asset jump diffusion models for option valuation. *Applied Numerical Mathematics*, 58(6), 743–782.
- Constantinides, G. M. (1990). Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy*, 98(3), 519–543.
- Cont, R., & Tankov, P. (2003). *Financial modelling with jump processes*. Chapman; Hall/CRC.

- Cox, J. C., & Huang, C.-f. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory*, 49(1), 33–83.
- Cvitanić, J., Lazrak, A., Martellini, L., & Zapatero, F. (2006). Dynamic portfolio choice with parameter uncertainty and the economic value of analysts’ recommendations. *The Review of Financial Studies*, 19(4), 1113–1156.
- Dai, M., Dong, Y., & Jia, Y. (2023). Learning equilibrium mean-variance strategy. *Mathematical Finance*, 33(4), 1166–1212.
- Dai, M., Sun, Y., Xu, Z. Q., & Zhou, X. Y. (2024). Learning to optimally stop diffusion processes, with financial applications. *arXiv preprint arXiv:2408.09242*.
- Das, S. R. (2002). The surprise element: Jumps in interest rates. *Journal of Econometrics*, 106(1), 27–65.
- Denkert, R., Pham, H., & Warin, X. (2024). Control randomisation approach for policy gradient and application to reinforcement learning in optimal switching. *arXiv preprint arXiv:2404.17939*.
- Deshpande, S., Walambe, R., Kotecha, K., Selvachandran, G., & Abraham, A. (2025). Advances and applications in inverse reinforcement learning: A comprehensive review. *Neural Computing and Applications*, 1–53.
- Detemple, J. B., & Zapatero, F. (1992). Optimal consumption-portfolio policies with habit formation. *Mathematical Finance*, 2(4), 251–274.
- Detemple, J. B., & Karatzas, I. (2003). Non-addictive habits: Optimal consumption-portfolio policies. *Journal of Economic Theory*, 113(2), 265–285.
- Dianetti, J., Ferrari, G., & Xu, R. (2024). Exploratory optimal stopping: A singular control formulation. *arXiv preprint arXiv:2408.09335*.
- Dong, Y. (2024). Randomized optimal stopping problem in continuous time and reinforcement learning algorithm. *SIAM Journal on Control and Optimization*, 62(3), 1590–1614.
- Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, 12(1), 219–245.
- Ebert, S., Wei, W., & Zhou, X. Y. (2020). Weighted discounting—on group diversity, time-inconsistency, and consequences for investment. *Journal of Economic Theory*, 189, 105089.
- Ekeland, I., & Pirvu, T. A. (2008). Investment and consumption without commitment. *Mathematics and Financial Economics*, 2(1), 57–86.
- Epstein, L. G., & Hynes, J. A. (1983). The rate of time preference and dynamic economic analysis. *Journal of Political Economy*, 91(4), 611–635.
- Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science*, 4(3), 282–289.

- Forsyth, P. A., & Vetzal, K. R. (2002). Quadratic convergence for valuing American options using a penalty method. *SIAM Journal on Scientific Computing*, 23(6), 2095–2122.
- Friedman, A. (1975). *Stochastic differential equations and applications, Volume II*. Springer.
- Fulop, A. (2011). Filtering methods. *Handbook of Computational Finance*, 439–467.
- Gao, X., Li, L., & Zhou, X. Y. (2024). Reinforcement learning for jump-diffusions. *arXiv preprint arXiv:2405.16449*.
- Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6), 359–483.
- Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory*, 118(2), 133–173.
- Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2), 141–153.
- Gullapalli, V. (1990). A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks*, 3(6), 671–692.
- Hambly, B., Xu, R., & Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3), 437–503.
- Harris, C., & Laibson, D. (2013). Instantaneous gratification. *The Quarterly Journal of Economics*, 128(1), 205–248.
- He, C., Kennedy, J. S., Coleman, T. F., Forsyth, P. A., Li, Y., & Vetzal, K. R. (2006). Calibration and hedging under jump diffusion. *Review of Derivatives Research*, 9, 1–35.
- Hsiau, S.-R., & Yang, J.-R. (2000). A natural variation of the standard secretary problem. *Statistica Sinica*, 10(2), 639–646.
- Hu, Y., Jin, H., & Zhou, X. Y. (2012). Time-inconsistent stochastic linear-quadratic control. *SIAM Journal on Control and Optimization*, 50(3), 1548–1572.
- Hu, Y., Jin, H., & Zhou, X. Y. (2017). Time-inconsistent stochastic linear-quadratic control: Characterization and uniqueness of equilibrium. *SIAM Journal on Control and Optimization*, 55(2), 1261–1279.
- Hull, J. C., & Basu, S. (2016). *Options, futures, and other derivatives*. Pearson Education India.
- Jia, Y., & Zhou, X. Y. (2025). Erratum to “q-learning in continuous time”. *Journal of Machine Learning Research*.
- Jia, Y., & Zhou, X. Y. (2022a). Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154), 1–55. <http://jmlr.org/papers/v23/21-0947.html>

- Jia, Y., & Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275), 1–50.
- Jia, Y., & Zhou, X. Y. (2023). q-learning in continuous time. *Journal of Machine Learning Research*, 24(161), 1–61.
- Jiang, R., Saunders, D., & Weng, C. (2022). The reinforcement learning Kelly strategy. *Quantitative Finance*, 22(8), 1445–1464.
- Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.
- Jorion, P. (1992). Portfolio optimization in practice. *Financial Analysts Journal*, 48(1), 68–74.
- Kalayci, C. B., Ertenlice, O., & Akbay, M. A. (2019). A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. *Expert Systems with Applications*, 125, 345–368.
- Karatzas, I., Lehoczky, J. P., Sethi, S. P., & Shreve, S. E. (1986). Explicit solution of a general consumption/investment problem. *Mathematics of Operations Research*, 11(2), 261–294.
- Karatzas, I., Lehoczky, J. P., & Shreve, S. E. (1987a). Optimal consumption and investment strategies in the infinite horizon. *Journal of Economic Theory*, 42(2), 297–313.
- Karatzas, I., Lehoczky, J. P., & Shreve, S. E. (1987b). Optimal portfolio and consumption decisions for a small investor on a finite horizon. *SIAM Journal on Control and Optimization*, 25(6), 1557–1586.
- Karatzas, I., & Zhao, X. (2001). Bayesian adaptive portfolio optimization. *Option pricing, interest rates and risk management*, 632–669.
- Karp, L. (2007). Non-constant discounting in continuous time. *Journal of Economic Theory*, 132(1), 557–568.
- Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, 73(6), 1849–1892.
- Kloeden, P. E., & Platen, E. (2018). Numerical methods for stochastic differential equations. In *Nonlinear dynamics and stochastic mechanics* (pp. 437–461). CRC Press.
- Kraft, H., Seifried, F. T., & Steffensen, M. (2013). Consumption-portfolio optimization with recursive utility in incomplete markets. *Finance and Stochastics*, 17, 161–196.
- Krusell, P., & Smith, A. A. (2003). Consumption-savings decisions with quasi-geometric discounting. *Econometrica*, 71(1), 365–375.
- Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443–477.
- Lesmana, N. S., & Pun, C. S. (2021). A subgame perfect equilibrium reinforcement learning approach to time-inconsistent problems. *arXiv preprint arXiv:2110.14295*.

- Lindley, D. V. (1961). Dynamic programming and decision theory. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 10(1), 39–51.
- Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2), 573–597.
- Luenberger, D. G. (2013). *Investment science* (Second). Oxford University Press.
- Maccheroni, F., Marinacci, M., & Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6), 1447–1498.
- Maheu, J. M., & McCurdy, T. H. (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics*, 18(1), 100–112.
- Marinacci, M. (2002). Probabilistic sophistication and multiple priors. *Econometrica*, 70(2), 755–764.
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91.
- McKean Jr., H. (1965). A free boundary problem for the heat equation arising from a problem in mathematical economics. *Industrial Management Review*, 6, 32–39.
- Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, 247–257.
- Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3(4), 373–413.
- Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2), 125–144.
- Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.
- Mina, K. F., Cheang, G. H., & Chiarella, C. (2015). Approximate hedging of options under jump-diffusion processes. *International Journal of Theoretical and Applied Finance*, 18(4), 1550024.
- Montgomery, H. L., & Vaughan, R. C. (2007). *Multiplicative number theory i: Classical theory*. Cambridge University Press.
- Mordecki, E. (2002). Optimal stopping and perpetual options for Lévy processes. *Finance and Stochastics*, 6, 473–493.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11), 2074–2102.
- Munk, C. (2008). Portfolio and consumption choice with stochastic investment opportunities and habit formation in preferences. *Journal of Economic Dynamics and Control*, 32(11), 3560–3589.
- Myers, S. C. (1968). A time-state-preference model of security valuation. *Journal of Financial and Quantitative Analysis*, 3(1), 1–33.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4), 308–313.

- Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. *Proceedings of the 17th International Conference on Machine Learning*, 663–670.
- Øksendal, B. K., & Sulem, A. (2005). *Applied stochastic control of jump diffusions*. Berlin: Springer.
- Park, K., Ahn, C. M., & Fujihara, R. (1993). Optimal hedged portfolios: The case of jump-diffusion risks. *Journal of International Money and Finance*, 12(5), 493–510.
- Peng, Y., Wei, P., & Wei, W. (2024). Deep penalty methods: A class of deep learning algorithms for solving high dimensional optimal stopping problems. *arXiv preprint arXiv:2405.11392*.
- Peskir, G., & Shiryaev, A. (2006). *Optimal stopping and free-boundary problems*. Springer.
- Pham, H. (1997). Optimal stopping, free boundary, and american option in a jump-diffusion model. *Applied Mathematics and Optimization*, 35, 145–164.
- Pham, H. (2011). Portfolio optimization under partial observation: Theoretical and numerical aspects. *Handbook of Nonlinear Filtering*.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd). Cambridge University Press.
- Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409), 2015–2020.
- Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, 23(3), 165–180.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3, 9–44.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd). MIT press.
- Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.
- Toivanen, J. (2008). Numerical valuation of european and american options under Kou’s jump-diffusion model. *SIAM Journal on Scientific Computing*, 30(4), 1949–1970.
- van Bilsen, S., Bovenberg, A. L., & Laeven, R. J. (2020). Consumption and portfolio choice under internal multiplicative habit formation. *Journal of Financial and Quantitative Analysis*, 55(7), 2334–2371.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269. <https://doi.org/10.1109/TIT.1967.1054010>
- Vlassis, N., Ghavamzadeh, M., Mannor, S., & Poupart, P. (2012). Bayesian reinforcement learning. *Reinforcement Learning: State-of-the-Art*, 359–386.

- Wang, H., Zariphopoulou, T., & Zhou, X. Y. (2020). Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(1), 8145–8178.
- Wang, H., & Zhou, X. Y. (2020). Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4), 1273–1308.
- Wang, Y., Liu, J., & Siu, T. K. (2024). Investment-consumption-insurance optimisation problem with multiple habit formation and non-exponential discounting. *Finance and Stochastics*, 28(1), 161–214.
- Weitzman, M. L. (2001). Gamma discounting. *American Economic Review*, 91(1), 260–271.
- Wen, W., Yuan, Y., & Yang, J. (2021). Reinforcement learning for options trading. *Applied Sciences*, 11(23), 11208.
- Wu, B., & Li, L. (2024). Reinforcement learning for continuous-time mean-variance portfolio selection in a regime-switching market. *Journal of Economic Dynamics and Control*, 158, 104787.
- Wu, H., & Li, Z. (2011). Multi-period mean-variance portfolio selection with markov regime switching and uncertain time-horizon. *Journal of Systems Science and Complexity*, 24(1), 140–155.
- Wu, H., Zeng, Y., & Yao, H. (2014). Multi-period markowitz’s mean–variance portfolio selection with state-dependent exit probability. *Economic modelling*, 36, 69–78.
- Xie, S., Li, Z., & Wang, S. (2008). Continuous-time portfolio selection with liability: Mean–variance model and stochastic LQ approach. *Insurance: Mathematics and Economics*, 42(3), 943–953.
- Xu, W., Wu, C., Xu, W., & Li, H. (2009). A jump-diffusion model for option pricing under fuzzy environments. *Insurance: Mathematics and Economics*, 44(3), 337–344.
- Yin, G., & Zhou, X. Y. (2004). Markowitz’s mean-variance portfolio selection with regime switching: From discrete-time models to their continuous-time limits. *IEEE Transactions on automatic control*, 49(3), 349–360.
- Yong, J. (2011). A deterministic linear quadratic time-inconsistent optimal control problem.
- Yong, J. (2012). Time-inconsistent optimal control problems and the equilibrium HJB equation. *arXiv preprint arXiv:1204.0568*.
- Yu, X. (2017). Optimal consumption under habit formation in markets with transaction costs and random endowments. *Annals of Applied Probability*, 27(2), 960–1002.
- Zhang, Y., Li, X., & Guo, S. (2018). Portfolio selection problems with markowitz’s mean–variance framework: A review of literature. *Fuzzy Optimization and Decision Making*, 17, 125–158.

- Zhao, Q., Shen, Y., & Wei, J. (2014). Consumption-investment strategies with non-exponential discounting and logarithmic utility. *European Journal of Operational Research*, 238(3), 824–835.
- Zhou, X. Y., & Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, 42, 19–33.
- Zhou, X. Y., & Yin, G. (2003). Markowitz’s mean-variance portfolio selection with regime switching: A continuous-time model. *SIAM Journal on Control and Optimization*, 42(4), 1466–1482.
- Zou, Z., Chen, S., & Wedge, L. (2014). Finite horizon consumption and portfolio decisions with stochastic hyperbolic discounting. *Journal of Mathematical Economics*, 52, 70–80.

Appendix A

APPENDICES

A.1 Proofs and Derivations of Chapter 2

A.1.1 Derivation of (2.14)

We herein derive (2.14) from (2.3), inspired by the exploratory formulation arguments in (H. Wang et al., 2020). Recall the original portfolio value dynamics without the exploratory extension (2.3)

$$dX_t^u = [r(t, \alpha_t)X_t^u + \rho(t, \alpha_t)\sigma(t, \alpha_t)u_t] dt + \sigma(t, \alpha_t)u_t dW_t.$$

With the exploratory extension, we know that each u_t at time t is sampled from the policy distribution $\pi_{t, \alpha_t}(\cdot)$ given the regime α_t . Consider simulating a path of portfolio values $\{X_t\}_{t \in [0, T]}$ and a path of investment controls $\{u_t\}_{t \in [0, T]}$. Then for any time $t \in [0, T]$ and a small time interval Δt ,

$$\Delta X_t = X_{t+\Delta t} - X_t \approx [r(t, \alpha_t)X_t + \rho(t, \alpha_t)\sigma(t, \alpha_t)u_t] \Delta t + \sigma(t, \alpha_t)u_t(W_{t+\Delta t} - W_t).$$

If we repetitively simulate N paths (denoting the j -th path with a superscript j) and compute

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N \Delta X_t^j &\approx \frac{1}{N} \sum_{j=1}^N [r(t, \alpha_t)X_t^j + \rho(t, \alpha_t)\sigma(t, \alpha_t)u_t^j] \Delta t + \frac{1}{N} \sum_{j=1}^N \sigma(t, \alpha_t)u_t^j(W_{t+\Delta t}^j - W_t^j) \\ &\xrightarrow{a.s.} \mathbb{E} \left[\int_{\mathcal{U}} [r(t, \alpha_t)X_t^\pi + \rho(t, \alpha_t)\sigma(t, \alpha_t)u] \pi_{t, \alpha_t}(u) du \Delta t \right] \end{aligned}$$

$$\begin{aligned}
& + \mathbb{E} \left[\int_{\mathcal{U}} \sigma(t, \alpha_t) u \pi_{t, \alpha_t}(u) du \right] \mathbb{E} [W_{t+\Delta t}^j - W_t^j] \\
& = \mathbb{E} \left[r(t, \alpha_t) X_t^\pi + \int_{\mathcal{U}} \rho(t, \alpha_t) \sigma(t, \alpha_t) u \pi_{t, \alpha_t}(u) du \right] \Delta t \\
& \frac{1}{N} \sum_{j=1}^N (\Delta X_t^j)^2 \approx \frac{1}{N} \sum_{j=1}^N (\sigma(t, \alpha_t) u_t^j)^2 \Delta t \xrightarrow{a.s.} \mathbb{E} \left[\int_{\mathcal{U}} \sigma^2(t, \alpha_t) u^2 \pi_{t, \alpha_t}(u) du \right] \Delta t,
\end{aligned}$$

where the almost sure convergence, $\xrightarrow{a.s.}$, follows from the Law of Large Numbers. Moreover, since the sample paths of portfolio values are randomly and independently simulated from its distribution X^π , the Law of Large Numbers again implies

$$\begin{aligned}
& \frac{1}{N} \sum_{i=1}^N \Delta X_t^i \xrightarrow{a.s.} \mathbb{E} [\Delta X_t^\pi], \\
& \frac{1}{N} \sum_{i=1}^N (\Delta X_t^i)^2 \xrightarrow{a.s.} \mathbb{E} [(\Delta X_t^\pi)^2].
\end{aligned}$$

This motivates us to formulate the exploratory version of the portfolio value process

$$\begin{aligned}
dX_t^\pi & = \left[r(t, \alpha_t) X_t^\pi + \int_{\mathcal{A}} \rho(t, \alpha_t) \sigma(t, \alpha_t) \cdot u \cdot \pi_{t, \alpha_t}(u) du \right] dt \\
& + \left(\sqrt{\int_{\mathcal{A}} \sigma^2(t, \alpha_t) u^2 \cdot \pi_{t, \alpha_t}(u) du} \right) dW_t.
\end{aligned}$$

A.1.2 Derivation of (2.20)

By Bellman's Principle of Optimality, we can derive the recursive form of the optimal value function, for $0 \leq t \leq s \leq T, x \in \mathbb{R}, i \in \{1, \dots, l\}$:

$$V^*(t, x, i) = \inf_{\pi \in \mathcal{A}^\pi} \mathbb{E} \left[V^\pi(s, X_s^\pi, \alpha_s) + \xi \int_t^s \int_{\mathcal{A}} \pi_{k, \alpha_k}(u) \log \pi_{k, \alpha_k}(u) du dk \middle| X_t^\pi = x, \alpha_t = i \right]. \quad (\text{A.1})$$

Following DPP arguments, we know by assuming V is smooth and applying Itô's formula to (A.1) that the optimal value function V^* satisfies the HJB equation

$$\begin{aligned} v_t(t, x, i) + v_x(t, x, i)r(t, i)x + \sum_{j=1}^l q_{ij}v(t, x, j) \\ + \min_{\pi_{t,i}(\cdot) \in \mathcal{A}^\pi} \left\{ \int_{\mathcal{A}} \left[\frac{1}{2} v_{xx}(t, x, i) \sigma^2(t, i) u^2 + v_x(t, x, i) \rho(t, i) \sigma(t, i) u \right. \right. \\ \left. \left. + \xi \log \pi(u|i) \right] \pi_{t,i}(u) du \right\} = 0. \end{aligned} \quad (\text{A.2})$$

Solving the minimization part in (A.2) yields the optimal policy distribution π^* , which is a Gaussian distribution with the mean coinciding with the optimal control of the MVRS problem (2.9).

$$\begin{aligned} \pi_{t,i}^*(u) &= \frac{\exp \left(-\frac{1}{\xi} \left[\frac{1}{2} v_{xx}(t, x, i) \sigma^2(t, i) u^2 + v_x(t, x, i) \rho(t, i) \sigma(t, i) u \right] \right)}{\int_{\mathcal{A}} \exp \left(-\frac{1}{\xi} \left[\frac{1}{2} v_{xx}(t, x, i) \sigma^2(t, i) u^2 + v_x(t, x, i) \rho(t, i) \sigma(t, i) u \right] \right) du} \\ &= N \left(u \left| -\frac{\rho(t, i) v_x(t, x, i)}{\sigma(t, i) v_{xx}(t, x, i)}, \frac{\xi}{\sigma^2(t, i) v_{xx}(t, x, i)} \right. \right). \end{aligned} \quad (\text{A.3})$$

Substituting π^* into (A.2) reduces the HJB equation to (2.20).

A.1.3 Proof of Theorem 2.3.1

We first state and prove the following lemma, which will be used in the proof of Theorem 2.3.1.

Lemma A.1.3.1 *Let $f(t, i) : [0, T] \times \{1, \dots, l\} \mapsto \mathbb{R}$ and $G(t, i) : [0, T] \times \{1, \dots, l\} \mapsto \mathbb{R}$ be two deterministic functions such that, for any regime $i \in \{1, \dots, l\}$,*

- (i) $f(\cdot, i)$ is continuous in time $t \in [0, T]$;
- (ii) $G(\cdot, i) \in \mathbb{C}^1([0, T])$, i.e., $G(t, i)$ is continuously differentiable with respect to $t \in [0, T]$.

We further suppose $f(t, i)$ and $G(t, i)$ also satisfy the following ODE with a terminal condition:

$$\begin{cases} \dot{G}(t, i) = -f(t, i) - \sum_{j=1}^l q_{ij}G(t, j) \\ G(T, i) = 0 \end{cases}$$

where q_{ij} is the (i, j) -entry of the Markov Chain generator matrix for the regime. Then, for $t \in [0, T]$ and regime $\alpha_t = i \in \{1, \dots, l\}$, it can be solved by $G(t, i) = \mathbb{E} \left[\int_t^T f(s, \alpha_s) ds \middle| \alpha_t = i \right]$, where the Markovian regime $\{\alpha_s \in \{1, \dots, l\}\}_{s \in [t, T]}$ is defined in Section 2.2.

For $t \in [0, T]$ and $\alpha_t = i \in \{1, \dots, l\}$, let $G(t, i) = \mathbb{E} \left[\int_t^T f(s, \alpha_s) ds \middle| \alpha_t = i \right]$. We verify that it satisfies the ODE system in Lemma A.1.3.1. First, the boundary condition is trivially satisfied. To verify the first equation of the ODE, we notice that $\dot{G}$ can be expressed as a limit:

$$\dot{G}(t, i) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} [G(t + \Delta t, \alpha_{t+\Delta t}) - G(t, i)],$$

where Δt is a small time interval. Moreover, by the tower property of conditional expectation, the right hand side becomes

$$\begin{aligned} & G(t + \Delta t, \alpha_{t+\Delta t}) - G(t, i) \\ &= G(t + \Delta t, \alpha_{t+\Delta t}) - \mathbb{E} [G(t + \Delta t, \alpha_{t+\Delta t}) | \alpha_t = i] + \mathbb{E} [G(t + \Delta t, \alpha_{t+\Delta t}) | \alpha_t = i] - G(t, i) \\ &= G(t + \Delta t, \alpha_{t+\Delta t}) - \mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_t = i \right] \\ &+ \mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_t = i \right] - \mathbb{E} \left[\int_t^T f(s, \alpha_s) ds \middle| \alpha_t = i \right] \\ &= G(t + \Delta t, \alpha_{t+\Delta t}) - \mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_t = i \right] + \mathbb{E} \left[- \int_t^{t+\Delta t} f(s, \alpha_s) ds \middle| \alpha_t = i \right]. \end{aligned}$$

Again by the tower property of conditional expectation,

$$\begin{aligned} \mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_t = i \right] &= \mathbb{E} \left[\mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_{t+\Delta t} \right] \middle| \alpha_t = i \right] \\ &= \sum_{j=1}^l p_{ij}(\Delta t) \mathbb{E} \left[\int_{t+\Delta t}^T f(s, \alpha_s) ds \middle| \alpha_{t+\Delta t} = j \right] \\ &= \sum_{j=1}^l p_{ij}(\Delta t) G(t + \Delta t, j), \end{aligned}$$

where $p_{ij}(\Delta t) = \mathbb{P}(\alpha_{t+\Delta t} = j | \alpha_t = i)$. Therefore,

$$\begin{aligned} & G(t + \Delta t, \alpha_{t+\Delta t}) - G(t, i) \\ &= - \sum_{j=1}^l p_{ij}(\Delta t) [G(t + \Delta t, j) - G(t + \Delta t, \alpha_{t+\Delta t})] - \mathbb{E} \left[\int_t^{t+\Delta t} f(s, \alpha_s) ds \middle| \alpha_t = i \right], \end{aligned}$$

as $\sum_{j=1}^l p_{ij}(\Delta t) = 1$. Finally, divide both sides by Δt and take the limit of $\Delta t \rightarrow 0+$. The left hand side becomes $\dot{G}$. On the right hand side, when $i \neq j$,

$$\lim_{\Delta t \rightarrow 0+} \frac{p_{ij}(\Delta t)}{\Delta t} = q_{ij}$$

and when $i = j$, $\lim_{\Delta t \rightarrow 0+} G(t + \Delta t, j) - G(t + \Delta t, \alpha_{t+\Delta t}) = 0$. This yields:

$$\begin{aligned} \dot{G}(t, i) &= -f(t, i) - \sum_{j=1}^l q_{ij} [G(t, j) - G(t, i)] \\ &= -f(t, i) - \sum_{j=1}^l q_{ij} G(t, j), \end{aligned}$$

where the last equation is given by $\sum_{j=1}^l q_{ij} = 0$.

Now, we are ready to prove Theorem 2.3.1.

The optimal policy distribution (2.21) can be easily derived from the optimal value function (2.22), following (A.3). Moreover, we notice that the optimal value function can be rewritten as a quadratic function of $(\lambda - z)$:

$$\begin{aligned} V^*(t, x, i) &= [P(t, i)H(t, i)^2 + C(t, i) - 1](\lambda - z)^2 \\ &\quad + 2[P(t, i)H(0, i)x - z](\lambda - z) + P(t, i)x^2 + D(t, i) - z^2. \end{aligned}$$

This yields the optimal Lagrange multiplier at initialization $t = 0$, which is the minimizer of $V^*(0, x_0, i_0)$, for some initial wealth $x_0 > 0$ and initial regime $i_0 \in \{1, \dots, l\}$.

The remainder of the proof is to verify that (2.22) solves the reduced HJB equation (2.20). We note that the partial derivatives of V^* are

$$\begin{aligned} V_t^*(t, x, i) &= \dot{P}(t, i)[x + (\lambda - z)H(t, i)]^2 + 2(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)] \cdot \dot{H}(t, i) \\ &\quad + (\lambda - z)^2 \dot{C}(t, i) + \dot{D}(t, i), \end{aligned}$$

$$\begin{aligned} V_x^*(t, x, i) &= 2P(t, i)[x + (\lambda - z)H(t, i)], \\ V_{xx}^*(t, x, i) &= 2P(t, i). \end{aligned}$$

Substituting these into the left hand side of (2.20) yields

$$\begin{aligned} & \dot{P}(t, i)[x + (\lambda - z)H(t, i)]^2 + 2(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)] \cdot \dot{H}(t, i) + (\lambda - z)^2 \dot{C}(t, i) + \dot{D}(t, i) \\ & + \sum_{j=1}^l q_{ij} \{P(t, j)[x + (\lambda - z)H(t, j)]^2 + (\lambda - z)^2 C(t, j) + D(t, j) - \lambda^2\} \\ & + 2r(t, i)xP(t, i)[x + (\lambda - z)H(t, i)] \\ & - \rho^2(t, i)P(t, i)[x + (\lambda - z)H(t, i)]^2 - \frac{\xi}{2} \log \left(\frac{\pi \xi}{\sigma^2(t, i)P(t, i)} \right). \end{aligned} \tag{A.4}$$

Adding 3 lines has no effect, but helps isolating P, H, C, D :

$$\begin{aligned} & \dot{P}(t, i)[x + (\lambda - z)H(t, i)]^2 + 2(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)] \cdot \dot{H}(t, i) + (\lambda - z)^2 \dot{C}(t, i) + \dot{D}(t, i) \\ & + \sum_{j=1}^l q_{ij} \{P(t, j)[x + (\lambda - z)H(t, j)]^2 + (\lambda - z)^2 C(t, j) + D(t, j)\} \\ & + \sum_{j=1}^l q_{ij} P(t, j)[x + (\lambda - z)H(t, i)]^2 - \sum_{j=1}^l q_{ij} P(t, j)[x + (\lambda - z)H(t, i)]^2 \\ & + \sum_{j=1}^l q_{ij} P(t, j) [2(\lambda - z)^2 H(t, i)(H(t, j) - H(t, i)) - 2(\lambda - z)^2 H(t, i)(H(t, j) - H(t, i))] \\ & + 2r(t, i)xP(t, i)[x + (\lambda - z)H(t, i)] \\ & + 2r(t, i)(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)]H(t, i) - 2r(t, i)(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)]H(t, i) \\ & - \rho^2(t, i)P(t, i)[x + (\lambda - z)H(t, i)]^2 - \frac{\xi}{2} \log \left(\frac{\pi \xi}{\sigma^2(t, i)P(t, i)} \right). \end{aligned} \tag{A.5}$$

Rearranging and grouping gives

$$\begin{aligned}
& \left\{ \dot{P}(t, i) - (\rho^2(t, i) - 2r(t, i))P(t, i) + \sum_{j=1}^l q_{ij}P(t, j) \right\} \cdot [x + (\lambda - z)H(t, i)]^2 \\
& + \left\{ \dot{H}(t, i) - r(t, i)H(t, i) + \frac{1}{P(t, i)} \sum_{j=1}^l q_{ij}P(t, j)(H(t, j) - H(t, i)) \right\} \\
& \quad \cdot 2(\lambda - z)P(t, i)[x + (\lambda - z)H(t, i)] \\
& + \left\{ \dot{C}(t, i) + \sum_{j=1}^l q_{ij} [P(t, j)(H(t, j) - H(t, i))^2 + C(t, j)] \right\} \cdot (\lambda - z)^2 \\
& + \left\{ \dot{D}(t, i) + \sum_{j=1}^l q_{ij}D(t, j) - \frac{\xi}{2} \log \left(\frac{\pi\xi}{\sigma^2(t, i)P(t, i)} \right) \right\} = 0,
\end{aligned} \tag{A.6}$$

indicating that the reduced HJB equation holds.

We remind the reader that the above equation holds because $P(t, i)$, $H(t, i)$, $C(t, i)$, $D(t, i)$ solves the following system of ODEs

$$\begin{cases} \dot{P}(t, i) = (\rho^2(t, i) - 2r(t, i))P(t, i) - \sum_{j=1}^l q_{ij}P(t, j) \\ P(T, i) = 1, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.7}$$

$$\begin{cases} \dot{H}(t, i) = r(t, i)H(t, i) - \frac{1}{P(t, i)} \sum_{j=1}^l q_{ij}P(t, j)(H(t, j) - H(t, i)) \\ H(T, i) = 1, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.8}$$

$$\begin{cases} \dot{C}(t, i) = - \sum_{j=1}^l q_{ij} [P(t, j)(H(t, j) - H(t, i))^2 + C(t, j)] \\ C(T, i) = 0, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.9}$$

$$\begin{cases} \dot{D}(t, i) = \frac{\xi}{2} \log \left(\frac{\pi\xi}{\sigma^2(t, i)P(t, i)} \right) - \sum_{j=1}^l q_{ij}D(t, j) \\ D(T, i) = 0, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.10}$$

(Zhou & Yin, 2003) discussed the existence and uniqueness of the solutions to (2.12) and (2.13), which follows since they are linear with uniformly bounded coefficients. Applying Lemma A.1.3.1, we can explicitly solve the ODEs for $C(t, i)$ and $D(t, i)$

$$\begin{aligned}
C(t, i) &= \mathbb{E} \left[\int_t^T \sum_{j=1}^l q_{\alpha_s j} P(s, j) (H(s, j) - H(s, \alpha_s))^2 ds \middle| \alpha_t = i \right] \\
&= \sum_{m=1}^l \sum_{j=1}^l \int_t^T p_{im}(s - t) q_{mj} P(s, j) (H(s, j) - H(s, m))^2 ds,
\end{aligned} \tag{A.11}$$

$$\begin{aligned}
D(t, i) &= -\mathbb{E} \left[\int_t^T \frac{\xi}{2} \log \left(\frac{\xi \pi}{\sigma(s, \alpha_s) P(s, \alpha_s)} \right) ds \middle| \alpha_t = i \right] \\
&= -\sum_{m=1}^l \int_t^T p_{im}(s-t) \frac{\xi}{2} \log \left(\frac{\pi \xi}{\sigma^2(s, m) P(s, m)} \right) ds.
\end{aligned} \tag{A.12}$$

Lastly, we notice that V^* is indeed smooth as all of its components P, H, C, D are smooth functions.

A.1.4 Proof of Corollary 2.3.1

By Theorem 2.3.1, the EMV problem 2.28 has an optimal value function taking the same form as (2.22):

$$V^*(t, x) = P(t)[x + (\lambda - z)H(t)]^2 + (\lambda - z)^2 C(t) + D(t) - \lambda^2, \tag{A.13}$$

where $P(t, i), H(t, i), C(t, i), D(t, i)$ solve the following system of ODEs, which are simplified from (2.12), (2.13), (2.26) and (2.27) to:

$$\begin{cases} \dot{P}(t) = (\rho^2 - 2r)P(t) \\ P(T) = 1, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.14}$$

$$\begin{cases} \dot{H}(t) = r(t)H(t) \\ H(T) = 1, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.15}$$

$$\begin{cases} \dot{C}(t) = 0 \\ C(T) = 0, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.16}$$

$$\begin{cases} \dot{D}(t) = \frac{\xi}{2} \log \left(\frac{\xi \pi}{\sigma^2 P(t)} \right) \\ D(T) = 0, \text{ for } i \in \{1, \dots, l\} \end{cases} \tag{A.17}$$

because $\{q_{ij}\}_{i,j=1,\dots,l} = 0$ when there is no regime switching. These ODEs can be easily solved:

$$P(t) = e^{-(\rho^2 - 2r)(T-t)}, \tag{A.18}$$

$$H(t) = e^{-r(T-t)}, \tag{A.19}$$

$$C(t) = 0, \tag{A.20}$$

$$D(t) = \frac{\xi(\rho^2 - 2r)}{4}(T^2 - t^2) - \frac{\xi}{2} \left[(\rho^2 - 2r)T - \log \frac{\sigma^2}{\pi \xi} \right] (T - t). \tag{A.21}$$

The optimal policy distribution can be easily derived through (2.31), and λ^* follows from (2.25) in Theorem 2.3.1.

A.1.5 Proof of Theorem 2.3.2

Consider $0 < t < s < T$, $X_t^\pi = x$, $\alpha_t = i$ and suppose V^π is smooth. Let $\{X_k^{\pi^*, t, x}\}_{k \in [t, T]}$ be the wealth process that follows policy π^* and starts with wealth x at time t . Then, by Itô's Formula,

$$\begin{aligned}
& V^\pi(s, X_s^{\pi^*, t, x}, \alpha_s) - V^\pi(t, x, i) \\
&= \int_t^s \left[V_t^\pi(k, X_k^{\pi^*, t, x}, \alpha_k) + V_x^\pi(k, X_k^{\pi^*, t, x}, \alpha_k) r(k, \alpha_k) X_k^{\pi^*, t, x} + \sum_{j=1}^l q_{ij} V^\pi(k, X_k^{\pi^*, t, x}, j) \right] dk \\
&+ \int_t^s \int_{\mathcal{A}} \left[\frac{1}{2} V_{xx}^\pi(k, X_k^{\pi^*, t, x}, \alpha_k) \sigma^2(k, \alpha_k) u^2 \right. \\
&\quad \left. + V_x^\pi(k, X_k^{\pi^*, t, x}, \alpha_k) \rho(k, \alpha_k) \sigma(k, \alpha_k) u \right] \pi_{k, \alpha_k}(u) du dk.
\end{aligned} \tag{A.22}$$

As an admissible policy π , we know that its corresponding *value function* V^π has the recursive form (A.1), by the DPP. Under the smoothness assumption on V^π , we apply Itô's Formula which yields the following HJB equation

$$\begin{aligned}
& V_t^\pi(t, x, i) + V_x^\pi(t, x, i) r(t, i) x + \sum_{j=1}^l q_{ij} V^\pi(t, x, j) \\
&+ \int_{\mathcal{A}} \left[\frac{1}{2} V_{xx}^\pi(t, x, i) \sigma^2(t, i) u^2 + V_x^\pi(t, x, i) \rho(t, i) \sigma(t, i) u + \xi \log \pi_{t, i}(u) \right] \pi_{t, i}(u) du = 0,
\end{aligned}$$

for $(t, x) \in [0, T] \times \mathbb{R}$ and $i \in \{1, \dots, l\}$. Moreover, since π^* is constructed from V^π through (2.31), and $\forall (t, x) \in [0, T] \times \mathbb{R}, i \in \{1, \dots, l\}$,

$$\begin{aligned}
\pi_{t, i}^*(\cdot) = \arg \min_{\pi_{t, i}(\cdot) \in \mathcal{A}^\pi} & \int_{\mathcal{A}} \left[\frac{1}{2} V_{xx}^\pi(t, x, i) \sigma^2(t, i) u^2 \right. \\
& \left. + V_x^\pi(t, x, i) \rho(t, i) \sigma(t, i) u + \xi \log \pi_{t, i}(u) \right] \pi_{t, i}(u) du.
\end{aligned}$$

Hence, $\forall(t, x) \in [0, T] \times \mathbb{R}, i \in \{1, \dots, l\}$,

$$\begin{aligned} & V_t^\pi(t, x, i) + V_x^\pi(t, x, i)r(t, i)x + \sum_{j=1}^l q_{ij}V^\pi(t, x, i) \\ & + \int_{\mathcal{A}} \left[\frac{1}{2}V_{xx}^\pi(t, x, i)\sigma^2(t, i)u^2 + V_x^\pi(t, x, i)\rho(t, i)\sigma(t, i)u + \xi \log \pi_{t,i}^*(u) \right] \pi_{t,i}^*(u)du \leq 0. \end{aligned}$$

Substituting this back into (A.22) yields

$$\begin{aligned} & V^\pi(s, X^{\pi^*, t, x}(s), \alpha_s) - V^\pi(t, x, i) \leq -\xi \int_t^s \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk \\ \implies & V^\pi(t, x, i) \geq V^\pi(s, X^{\pi^*, t, x}(s), \alpha_s) + \xi \int_t^s \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk. \end{aligned}$$

Setting $s = T$ and taking expectation on both sides conditioned on $X_t = x, \alpha_t = i$, we have:

$$\begin{aligned} & \mathbb{E}[V^\pi(t, x, i) | X_t = x, \alpha_t = i] \\ & \geq \mathbb{E} \left[V^\pi(T, X_T^{\pi^*, t, x}, \alpha_T) + \xi \int_t^T \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk \middle| X_t = x, \alpha_t = i \right] \\ & = \mathbb{E} \left[V^{\pi^*}(T, X_T^{\pi^*, t, x}, \alpha_T) + \xi \int_t^T \int_{\mathcal{A}} \pi_{k, \alpha_k}^*(u) \log \pi_{k, \alpha_k}^*(u) du dk \middle| X_t = x, \alpha_t = i \right] \\ & = V^{\pi^*}(t, x, i). \end{aligned}$$

The second to the last equation is because $V^\pi(T, x, i) = V^{\pi^*}(T, x, i) = (x + (\lambda - z))^2 - \lambda^2$, for all $x \in \mathbb{R}, i \in \{1, \dots, l\}$, while the last equation is implied by Bellman's Principle of Optimality. Finally, note that the left-hand-side is $V^\pi(t, x, i)$, which concludes the proof.

A.2 Proofs of Chapter 3

A.2.1 Proof of Lemma 3.2.0.1

Suppose $\pi \in \mathcal{A}$. By the fact that $\text{Var}(U) = E[U^2] - (E[U])^2 \geq 0$ for $U \sim \pi$,

$$\int_{\mathbb{R}} u^2 \pi_{t,x}(u) du \geq \left(\int_{\mathbb{R}} u \pi_{t,x}(u) du \right)^2.$$

Furthermore, because $u = G^{\pi_{t,x}}(v)$ for $(t, x) \in [0, T] \times \mathbb{R}$ and hence $dv = \pi_{t,x}(u)du$, we have

$$\int_0^1 |G^{\pi_{t,x}}(v)|^2 dv = \int_{\mathbb{R}} u^2 \pi_{t,x}(u) du.$$

Then, by (2) of Definition 3.2.1 and the conditions on the Lévy measure (3.4), we know there exists a constant C_4 such that

$$\left(\int_{\mathbb{R}} u \pi_{t,x}(u) du \right)^2 \leq C_4(1 + x^2)$$

$$\int_{\mathbb{R} \times [0,1]} |G^{\pi_{t,x}}(v)(e^z - 1)|^2 \nu(dz) dv = \left(\int_{\mathbb{R}} (e^z - 1)^2 \nu(dz) \right) \times \left(\int_0^1 |G^{\pi_{t,x}}(v)|^2 dv \right) \leq C_4(1 + x^2)$$

for any $(t, x) \in [0, T] \times \mathbb{R}$. Therefore, there exists $C' := \max(C_1, 2C_4)$ such that

$$\left(\int_{\mathbb{R}} u \pi_{t,x}(u) du \right)^2 + \int_{\mathbb{R}} u^2 \pi_{t,x}(u) du + \int_{\mathbb{R} \times [0,1]} |G^{\pi_{t,x}}(v)(e^z - 1)|^2 \nu(dz) du \leq C'(1 + x^2),$$

for any $(t, x) \in [0, T] \times \mathbb{R}$. Hence, the drift, diffusion and jump components of the SDE (3.8) have at most linear growth. Furthermore, by (3) of Definition 3.2.1, the drift, diffusion and jump components of the SDE (3.8) are also Lipschitz continuous. According to Øksendal and Sulem, 2005, there exists a unique càdlàg adapted solution X^π to the SDE (3.8) such that

$$\mathbb{E} \left[\sup_{t \in [0, T]} (X_t^\pi)^2 \right] < \infty.$$

A.2.2 Proof of Theorem 3.3.1

First, by Itô's formula, we have

$$\mathbb{E}_{t,x}[g(T, X_T^{\pi*})] = g(t, x) + \mathbb{E}_{t,x} \left[\int_t^T (\partial_t + \mathcal{L}^\pi) g(s, X_s^{\pi*}) ds \right]. \quad (\text{A.23})$$

Then, by (3.14) and (3.15), we get

$$g^{\pi*}(t, x) \equiv g(t, x; \pi^*) := \mathbb{E}_{t,x}(X_T^{\pi*}), \quad (\text{A.24})$$

which proves (3.17).

Next, we prove (3.16). Again by Itô's formula, (3.13) and (3.15),

$$\begin{aligned} V(t, x) &= \mathbb{E}_{t,x} \left[- \int_t^T (\partial_t + \mathcal{L}^\pi) V(s, X_s^{\pi^*}) ds + V(T, X_T^{\pi^*}) \right] \\ &= \mathbb{E}_{t,x} \left[\int_t^T \lambda H(\pi_s^*) + \gamma g^{\pi^*}(s, X_s^{\pi^*}) \mathcal{L}^\pi g^{\pi^*}(s, X_s^{\pi^*}) - \frac{\gamma}{2} \mathcal{L}^\pi (g^{\pi^*}(s, X_s^{\pi^*}))^2 ds \right] + \mathbb{E}_{t,x}[X_T^{\pi^*}]. \end{aligned}$$

Note that for some $\pi \in \mathcal{A}$, $(t, x) \in [0, T] \times \mathbb{R}$,

$$\frac{\gamma}{2} \partial_t (g^\pi(t, x))^2 = \gamma g^\pi(t, x) \partial_t g^\pi(t, x) = -\gamma g^\pi(t, x) \mathcal{L}^\pi g^\pi(t, x).$$

So,

$$\begin{aligned} V(t, x) &= \mathbb{E}_{t,x} \left[\int_t^T \lambda H(\pi_s^*) ds - \frac{\gamma}{2} \int_t^T (\partial_t + \mathcal{L}^\pi) (g^{\pi^*}(s, X_s^{\pi^*}))^2 ds \right] + \mathbb{E}_{t,x}[X_T^{\pi^*}] \\ &= \mathbb{E}_{t,x} \left[\int_t^T \lambda H(\pi_s^*) ds - \frac{\gamma}{2} (g^{\pi^*}(T, X_T^{\pi^*}))^2 + \frac{\gamma}{2} (g^{\pi^*}(t, x))^2 \right] + \mathbb{E}_{t,x}[X_T^{\pi^*}] \\ &= \mathbb{E}_{t,x} \left[\int_t^T \lambda H(\pi_s^*) ds - \frac{\gamma}{2} (X_T^{\pi^*})^2 + X_T^{\pi^*} \right] + \frac{\gamma}{2} (\mathbb{E}_{t,x}(X_T^{\pi^*}))^2 \\ &= J(t, x; \pi^*), \end{aligned}$$

where the last equation is by (3.9).

Lastly, we show that π^* , the maximizer of (3.13), is an equilibrium policy. Consider a perturbed policy as in Definition 3.3.1. That is, for a small time interval $\epsilon > 0$, $\pi^\epsilon = \{\pi_{s,y}^\epsilon(\cdot)\}_{(s,y) \in [0,T] \times \mathbb{R}}$, with

$$\pi_{s,y}^\epsilon(\cdot) = \begin{cases} \nu(\cdot), & (s, y) \in [t, t + \epsilon) \times \mathbb{R}, \\ \pi_{s,y}^*(\cdot), & (s, y) \in [t + \epsilon, T] \times \mathbb{R}, \end{cases}$$

where $\nu(\cdot) \in \mathcal{P}(\mathbb{R})$. Under π^ϵ ,

$$\begin{aligned} J(t, x; \pi^\epsilon) &= \mathbb{E}_{t,x} \left[\int_t^{t+\epsilon} \lambda H(\nu) ds + \int_{t+\epsilon}^T \lambda H(\pi_s) ds + X_T^{\pi^\epsilon} - \frac{\gamma}{2} (X_T^{\pi^\epsilon})^2 \right] + \frac{\gamma}{2} (\mathbb{E}_{t,x}(X_T^{\pi^\epsilon}))^2 \\ &= \epsilon \lambda H(\nu) + \mathbb{E}_{t,x}(J(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon}; \pi^\epsilon)) + \frac{\gamma}{2} (\mathbb{E}_{t,x}(X_T^{\pi^\epsilon}))^2 - \mathbb{E}_{t,x} \left[\frac{\gamma}{2} (\mathbb{E}_{t+\epsilon, X_{t+\epsilon}^{\pi^\epsilon}}(X_T^{\pi^\epsilon}))^2 \right] \end{aligned}$$

$$= \epsilon \lambda H(\nu) + \mathbb{E}_{t,x}(V(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon})) - \frac{\gamma}{2} \mathbb{E}_{t,x} [(g^{\pi^\epsilon}(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon}))^2 - (g^{\pi^\epsilon}(t, x))^2].$$

By Itô's formula, we have

$$\begin{aligned} \mathbb{E}_{t,x}(V(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon})) - V(t, x) &= \mathbb{E}_{t,x} \left[\int_t^{t+\epsilon} (\partial_t + \mathcal{L}^\nu) V(s, X_s^\nu) ds \right] \\ &= \epsilon (\partial_t + \mathcal{L}^\nu) V(t, x) + O(\epsilon), \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{t,x} [(g^{\pi^\epsilon}(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon}))^2] - (g^{\pi^\epsilon}(t, x))^2 &= \mathbb{E}_{t,x} \left[\int_t^{t+\epsilon} (\partial_t + \mathcal{L}^\nu) (g^\nu(s, X_s^\nu))^2 ds \right] \\ &= \epsilon (\partial_t + \mathcal{L}^\nu) (g^\nu(t, x))^2 + O(\epsilon) \\ &= \epsilon (-2g^\nu(t, x) \mathcal{L}^\nu g^\nu(t, x) + \mathcal{L}^\nu (g^\nu(t, x))^2) + O(\epsilon). \end{aligned}$$

Combining these together, we have

$$\begin{aligned} J(t, x; \pi^\epsilon) - V(t, x) &= J(t, x; \pi^\epsilon) - \mathbb{E}_{t,x}(V(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon})) + \mathbb{E}_{t,x}(V(t + \epsilon, X_{t+\epsilon}^{\pi^\epsilon})) - V(t, x) \\ &= \epsilon \left(\lambda H(\nu) - \frac{\gamma}{2} \mathcal{L}^\nu (g^\nu(t, x))^2 + \gamma g^\nu(t, x) \mathcal{L}^\nu g^\nu(t, x) + (\partial_t + \mathcal{L}^\nu) V(t, x) \right) + O(\epsilon). \end{aligned}$$

We conclude from (3.13) that

$$\limsup_{\epsilon \downarrow 0} \frac{J(t, x; \pi^\epsilon) - J(t, x; \pi^*)}{\epsilon} = \frac{J(t, x; \pi^\epsilon) - V(t, x)}{\epsilon} \leq 0,$$

which verifies that π^* is an equilibrium policy.

A.2.3 Proof of Theorem 3.3.2

According to the objective function (3.9) and the definition of the auxiliary value function (3.17), we conjecture that the value function is quadratic in x and the auxiliary value function is linear in x , i.e.,

$$V(t, x) = A(t)x^2 + B(t)x + C(t), \tag{A.25}$$

$$g^\pi(t, x) = g(t, x; \pi) = x + h^\pi(t), \tag{A.26}$$

with $A(T) = C(T) = h^\pi(T) = 0$ and $B(T) = 1$. Under this conjecture and by the infinitesimal generator (3.12), for $(t, x) \in [0, T] \times \mathbb{R}$,

$$\mathcal{L}^\pi g^\pi(t, x) = \mathcal{L}^\pi(x + h^\pi(t)) = \int_{\mathbb{R}} u(\mu - r)\pi(u)du,$$

and

$$\begin{aligned} \mathcal{L}^\pi(g^\pi(t, x))^2 &= \mathcal{L}^\pi(x^2 + (h^\pi(t))^2 + 2xh^\pi(t)) \\ &= \int_{\mathbb{R}} u(\mu - r)\pi(u)du \cdot (2x + 2h^\pi(t)) + \int_{\mathbb{R}} \sigma^2 u^2 \pi(u)du \\ &\quad + \int_{\mathbb{R}} \int_{\mathbb{R}} [(g^\pi)^2(t, x + u(e^z - 1)) - (g^\pi)^2(t, x) - u(e^z - 1)\partial_x(g^\pi)^2]\pi(u)du \nu(dz) \\ &= \int_{\mathbb{R}} u(\mu - r)\pi(u)du \cdot (2x + 2h^\pi(t)) + \int_{\mathbb{R}} \sigma^2 u^2 \pi(u)du + \int_{\mathbb{R}} u^2 \delta^2 \pi(u)du, \end{aligned}$$

where $\delta^2 := \int_{\mathbb{R}} (e^z - 1)^2 \nu(dz)$. The last equation is given by the fact that, for some function $\psi(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R})$ and ψ is quadratic in x , the second order Taylor expansion yields $\psi(t, x + \Delta x) = \psi(t, x) + \psi_x(t, x)\Delta x + \frac{1}{2}\psi_{xx}(t, x)(\Delta x)^2$, and so,

$$\begin{aligned} &\int_{\mathbb{R}} \int_{\mathbb{R}} [\psi(t, x + u(e^z - 1)) - \psi(t, x) - u(e^z - 1)\psi_x]\pi(u)du \nu(dz) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{2}u^2(e^z - 1)^2 \pi(u)du \nu(dz) \cdot \psi_{xx}. \end{aligned}$$

So, from (3.13) and by (3.12),

$$\begin{aligned} &(\partial_t + \mathcal{L}^\pi)V + \lambda H(\pi) + \gamma g^\pi \mathcal{L}^\pi g^\pi - \frac{\gamma}{2} \mathcal{L}^\pi(g^\pi)^2 \\ &= V_t + \int_{\mathbb{R}} \left[u(\mu - r)V_x + \frac{1}{2}u^2\sigma^2 V_{xx} + \int_{\mathbb{R}} (V(t, x + u(e^z - 1)) - V(t, x) - u(e^z - 1)V_x)\nu(dz) \right] \pi_t(u)du \\ &\quad - \int_{\mathbb{R}} \lambda \pi_t(u) \log \pi_t(u)du + \int_{\mathbb{R}} \gamma(x + h^\pi(t))u(\mu - r)\pi(u)du \\ &\quad - \frac{\gamma}{2} \left[\int_{\mathbb{R}} u(\mu - r)\pi_t(u)du \cdot (2x + 2h^\pi(t)) + \int_{\mathbb{R}} \sigma^2 u^2 \pi_t(u)du + \int_{\mathbb{R}} u^2 \delta^2 \pi_t(u)du \right] \\ &= V_t + \int_{\mathbb{R}} \left[u(\mu - r)V_x + \frac{V_{xx} - \gamma}{2}u^2(\sigma^2 + \delta^2) - \lambda \log \pi_t(u) \right] \pi_t(u)du. \end{aligned}$$

The maximizer π^* that attains the supremum of (3.13) is

$$\begin{aligned}\pi_t^*(u) &= \exp \left(\frac{1}{\lambda} \left[u(\mu - r)V_x + \frac{V_{xx} - \gamma}{2} u^2 (\sigma^2 + \delta^2) - 1 \right] \right) \\ &\propto \exp \left(\frac{1}{2} \frac{(\sigma^2 + \delta^2)(V_{xx} - \gamma)}{\lambda} \left[u^2 + \frac{2(\mu - r)V_x}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} u \right] \right) \\ &\sim N \left(-\frac{(\mu - r)V_x}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)}, -\frac{\lambda}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} \right).\end{aligned}\tag{A.27}$$

Substituting this back into (3.13), we have

$$\begin{aligned}V_t - \frac{(\mu - r)^2 V_x^2}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} + \frac{V_{xx} - \gamma}{2} (\sigma^2 + \delta^2) \left[-\frac{\lambda}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} + \frac{(\mu - r)^2 V_x^2}{(\sigma^2 + \delta^2)^2 (V_{xx} - \gamma)^2} \right] \\ + \frac{\lambda}{2} \log \left(-\frac{2\pi e \lambda}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} \right) \\ = V_t - \frac{(\mu - r)^2 V_x^2}{2(\sigma^2 + \delta^2)(V_{xx} - \gamma)} + \frac{\lambda}{2} \log \left(-\frac{2\pi \lambda}{(\sigma^2 + \delta^2)(V_{xx} - \gamma)} \right) = 0.\end{aligned}$$

Under the conjectured forms of the value function and auxiliary value function, the left-hand-side of the above equation becomes

$$\begin{aligned}&A_t(t)x^2 + B_t(t)x + C_t(t) - \frac{(\mu - r)^2 (2A(t)x + B(t))^2}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} + \frac{\lambda}{2} \log \left(-\frac{2\pi \lambda}{(\sigma^2 + \delta^2)(2A(t) - \gamma)} \right) \\ &= \left\{ A_t(t) - \frac{4(\mu - r)^2 (A(t))^2}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} \right\} x^2 + \left\{ B_t(t) - \frac{4(\mu - r)^2 A(t)B(t)}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} \right\} x \\ &\quad + \left\{ C_t(t) - \frac{(\mu - r)^2 (B(t))^2}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} + \frac{\lambda}{2} \log \left(-\frac{2\pi \lambda}{(\sigma^2 + \delta^2)(2A(t) - \gamma)} \right) \right\} = 0.\end{aligned}$$

So, A, B, C can be solved by 3 ODEs.

$$A_t(t) - \frac{4(\mu - r)^2 (A(t))^2}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} = 0,\tag{A.28}$$

$$B_t(t) - \frac{4(\mu - r)^2 A(t)B(t)}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} = 0,\tag{A.29}$$

$$C_t(t) - \frac{(\mu - r)^2 (B(t))^2}{2(\sigma^2 + \delta^2)(2A(t) - \gamma)} + \frac{\lambda}{2} \log \left(-\frac{2\pi \lambda}{(\sigma^2 + \delta^2)(2A(t) - \gamma)} \right) = 0.\tag{A.30}$$

Moreover, we know from (3.14) that

$$\begin{aligned} (\partial_t + \mathcal{L}^\pi)g^\pi(t, x) &= h_t^\pi(t) + \int_{\mathbb{R}} u(\mu - r)\pi_t(u)du \\ &= h_t^\pi(t) - \frac{(\mu - r)^2(2A(t)x + B(t))}{(\sigma^2 + \delta^2)(2A(t) - \gamma)} = 0. \end{aligned}$$

This tells us that $A(t) = 0, \forall t \in [0, T]$. So,

$$h_t^\pi(t) + \frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} = 0, \quad (\text{A.31})$$

which yields (3.20). Then, it follows from the above ODEs that $B(t) = 1, \forall t \in [0, T]$ and

$$C_t(t) + \frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) = 0, \quad (\text{A.32})$$

which yields (3.19). This indicates that the value function is in fact linear in x and is $V(t, x) = x + C(t)$. Substituting $V(t, x)$ into the equilibrium policy (A.27) yields

$$\pi_t^*(u) \sim N \left(\frac{(\mu - r)}{\gamma(\sigma^2 + \delta^2)}, \frac{\lambda}{\gamma(\sigma^2 + \delta^2)} \right), \quad (\text{A.33})$$

which is (3.18). Since $\pi_t^*(u)$ does not depend on the portfolio value x , it is easy to check that it is admissible according to Definition 3.2.1, which completes the proof.

A.2.4 Proof of Theorem 3.4.1

Consider a parameterized policy π^θ , its corresponding value function V^θ and auxiliary value function g^θ , and a process M^θ defined in (3.26). Then, by Itô's formula, (3.25) and (3.24),

$$\begin{aligned} dM_t^\theta &= dX_t^\theta + C_t^\theta(t)dt - \gamma g^\theta d(g^\theta) - \frac{\gamma}{2} d[g^\theta] + \lambda H(\pi_t^\theta)dt \\ &= \int_{\mathbb{R}} u(\mu_{true} - r)\pi_t^\theta(u)du dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \\ &\quad - \left(\frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right) dt - \gamma g^\theta [dX_t^\theta + h_t^\theta(t)dt] \\ &\quad - \frac{\gamma}{2} d[X_t^\theta] + \frac{\lambda}{2} \log \left(\frac{2\pi e\lambda}{\gamma(\sigma^2 + \delta^2)} \right) dt \end{aligned}$$

$$\begin{aligned}
&= \int_{\mathbb{R}} u(\mu_{true} - r)\pi_t^\theta(u) du dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \\
&\quad - \left(\frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right) dt \\
&\quad - \gamma g^\theta \left[\int_{\mathbb{R}} (\mu_{true} - r)\pi_t^\theta(u) du dt - \frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \right] \\
&\quad - \frac{\gamma}{2} \left[\int_{\mathbb{R}} \sigma_{true}^2 u^2 \pi_t(u) du dt + \int_{\mathbb{R}} \int_{\mathbb{R}} u^2 (e^z - 1)^2 \pi_t(u) du dN(dt, dz) \right] \\
&\quad + \frac{\lambda}{2} \log \left(\frac{2\pi e \lambda}{\gamma(\sigma^2 + \delta^2)} \right) dt \\
&= \int_{\mathbb{R}} u(\mu_{true} - r)\pi_t^\theta(u) du dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \\
&\quad - \left(\frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right) dt \\
&\quad - \gamma g^\theta \left[\int_{\mathbb{R}} (\mu_{true} - r)\pi_t^\theta(u) du dt - \frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \right] \\
&\quad - \frac{\gamma}{2} \left[\int_{\mathbb{R}} \sigma_{true}^2 u^2 \pi_t(u) du dt + \int_{\mathbb{R}} \int_{\mathbb{R}} u^2 (e^z - 1)^2 \pi_t(u) du \nu(dz) dt + (...) \tilde{N}(dt, dz) \right] \\
&\quad + \frac{\lambda}{2} \log \left(\frac{2\pi e \lambda}{\gamma(\sigma^2 + \delta^2)} \right) dt \\
&= \frac{(\mu_{true} - r)(\mu - r)}{\lambda(\sigma^2 + \delta^2)} dt + (...)dW_t + (...)d\tilde{N}(dt, dz) \\
&\quad - \left(\frac{(\mu - r)^2}{2\gamma(\sigma^2 + \delta^2)} + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\gamma(\sigma^2 + \delta^2)} \right) \right) dt \\
&\quad - \gamma g^\theta \left[\frac{(\mu_{true} - r)(\mu - r)}{\gamma(\sigma^2 + \delta^2)} - \frac{(\mu - r)^2}{\gamma(\sigma^2 + \delta^2)} \right] dt \\
&\quad - \frac{\gamma(\sigma_{true}^2 + \delta_{true}^2)}{2} \left(\frac{\lambda}{\gamma(\sigma^2 + \delta^2)} + \frac{(\mu - r)^2}{\gamma^2(\sigma^2 + \delta^2)^2} \right) dt + \frac{\lambda}{2} \log \left(\frac{2\pi e \lambda}{\gamma(\sigma^2 + \delta^2)} \right) dt.
\end{aligned}$$

The drift term is zero when $\theta = \theta_{true}$, indicating that M^θ is a martingale.

A.3 Proofs of Chapter 4

A.3.1 Proof of Proposition 4.2.1

We define the measures of decreasing impatience under β_t^{exp} and $\beta_t^{\text{non-exp}}$ as

$$P_t^{\text{exp}} := -\frac{(\log(\beta_t^{\text{exp}}))''}{(\log(\beta_t^{\text{exp}}))'} \quad \text{and} \quad P_t^{\text{non-exp}} := -\frac{(\log(\beta_t^{\text{non-exp}}))''}{(\log(\beta_t^{\text{non-exp}}))'},$$

respectively. By Proposition 1 of Ebert et al. (2020), we have, for any $t \in [0, T]$,

$$0 = P_t^{\text{exp}} \leq P_t^{\text{non-exp}}.$$

Further, by Proposition 2 and Theorem 2 of (Ebert et al., 2020), it follows that, for any $t \in [0, T]$,

$$\beta_t^{\text{exp}} \leq \beta_t^{\text{non-exp}}.$$

By Theorem 4.2.1, we conclude that, for any $t \in [0, T]$,

$$A^{\text{exp}}(t) := \beta_{T-t}^{\text{exp}} + \int_t^T \beta_{T-l}^{\text{exp}} dl \leq \beta_{T-t}^{\text{non-exp}} + \int_t^T \beta_{T-l}^{\text{non-exp}} dl := A^{\text{non-exp}}(t),$$

and

$$c^{*,\text{exp}}(t) = \frac{1}{A^{\text{exp}}(t)} \geq \frac{1}{A^{\text{non-exp}}(t)} = c^{*,\text{non-exp}}(t).$$

A.3.2 Proof of Theorem 4.4.1

By Itô's formula, we have that

$$\mathbb{E}_{t,x} [f(T, X_T^{\pi^*}, s)] = f(t, x, s) + \mathbb{E}_{t,x} \left[\int_t^T \mathcal{L}^{\pi^*} f(l, X_l^{\pi^*}, s) dl \right].$$

It follows from (4.14) and (4.13) that

$$f(t, x, s) = \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^{\pi^*} + \int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right],$$

which proves (4.16). Note that we use the simplified notation $\pi_t^*(u, c) := \pi_{t, X_t^{\pi^*}}^*(u, c)$.

Next we prove (4.15). By Itô's formula, (4.12), and (4.14)

$$\begin{aligned}
V(t, x) &= \mathbb{E}_{t,x} \left[- \int_t^T \mathcal{L}^{\pi^*} V(l, X_l^{\pi^*}) dl + V(T, X_T^{\pi^*}) \right] \\
&= \mathbb{E}_{t,x} \left[\int_t^T \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl - \int_t^T \frac{\partial f}{\partial s}(l, X_l^{\pi^*}, l) dl \right] \\
&\quad + \mathbb{E}_{t,x}[\log X_T^{\pi^*}].
\end{aligned} \tag{A.34}$$

We denote by $f(t, x) := f(t, x, t)$. By (4.11),

$$\begin{aligned}
\mathcal{L}^{\pi^*} f(t, x) &= \frac{\partial f}{\partial t}(t, x) + \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu - r)u - c) \pi_{t,x}^*(u, c) dc du \right) x \frac{\partial f}{\partial x}(t, x) \\
&\quad + \frac{1}{2} x^2 \left(\int_{\mathbb{R}} \int_{\mathbb{R}_+} \sigma^2 u^2 \pi_{t,x}^*(u, c) dc du \right) \frac{\partial^2 f}{\partial x^2}(t, x) \\
&= \frac{\partial f}{\partial t}(t, x, t) + \frac{\partial f}{\partial s}(t, x, t) + \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu - r)u - c) \pi_{t,x}^*(u, c) dc du \right) x \frac{\partial f}{\partial x}(t, x, t) \\
&\quad + \frac{1}{2} x^2 \left(\int_{\mathbb{R}} \int_{\mathbb{R}_+} \sigma^2 u^2 \pi_{t,x}^*(u, c) dc du \right) \frac{\partial^2 f}{\partial x^2}(t, x, t) \\
&= \mathcal{L}^{\pi^*} f(t, x, t) + \frac{\partial f}{\partial s}(t, x, t),
\end{aligned} \tag{A.35}$$

where the second equality is by the total derivative formula. By (A.35), (4.13) with $t = k = s$, and Itô's formula, it follows that

$$\begin{aligned}
&\mathbb{E}_{t,x} \left[- \int_t^T \frac{\partial f}{\partial s}(l, X_l^{\pi^*}, l) dl \right] \\
&= \mathbb{E}_{t,x} \left[\int_t^T \mathcal{L}^{\pi^*} f(l, X_l^{\pi^*}, l) dl \right] - \mathbb{E}_{t,x} \left[\int_t^T \mathcal{L}^{\pi^*} f(l, X_l^{\pi^*}) dl \right] \\
&= -\mathbb{E}_{t,x} \left[\int_t^T \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right] \\
&\quad - (\mathbb{E}_{t,x} [f(T, X_T^{\pi^*}, T)] - f(t, x, t)) \\
&= -\mathbb{E}_{t,x} \left[\int_t^T \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl + \log X_T^{\pi^*} \right] \\
&\quad + \mathbb{E}_{t,x} \left[\beta_{T-t} \log X_T^{\pi^*} + \int_t^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right],
\end{aligned}$$

where the last equality is by (4.16) and (4.14). Substituting the above expression of $\mathbb{E}_{t,x} \left[- \int_t^T \frac{\partial f}{\partial s}(l, X_l^{\pi^*}, l) dl \right]$ into (A.34) yields

$$\begin{aligned} V(t, x) &= \mathbb{E}_{t,x} \left[\beta_{T-t} \log X_T^{\pi^*} + \int_t^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right] \\ &= J^{\pi^*}(t, x). \end{aligned}$$

It remains to show that π^* , the maximizer of (4.12), is an equilibrium policy. Consider the perturbed policy defined in Definition 4.3.2, that is,

$$\pi_{l,y}^\varepsilon(\cdot) = \begin{cases} v(\cdot), & l \in [t, t + \varepsilon), \\ \pi_{l,y}^*(\cdot), & l \in [t + \varepsilon, T], \end{cases}$$

where $v(\cdot)$ is an arbitrary distribution on $\mathbb{R} \times \mathbb{R}_+$. Then

$$\begin{aligned} J^{\pi^\varepsilon}(t, x) &= \mathbb{E}_{t,x} \left[\beta_{T-t} \log X_T^{\pi^\varepsilon} + \int_t^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^\varepsilon}) - \lambda \log(\pi_l^\varepsilon(u, c))) \pi_l^\varepsilon(u, c) dc du dl \right] \\ &= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^v) - \lambda \log(v(u, c))) v(u, c) dc du dl + \beta_{T-t} \log X_T^{\pi^\varepsilon} \right] \\ &\quad + \mathbb{E}_{t,x} \left[\int_{t+\varepsilon}^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^\varepsilon}) - \lambda \log(\pi_{l,X_l^{\pi^\varepsilon}}^*(u, c))) \pi_{l,X_l^{\pi^\varepsilon}}^*(u, c) dc du dl \right], \end{aligned}$$

Moreover, the last two terms in the above equation can be simplified as

$$\begin{aligned} &\mathbb{E}_{t,x} \left[\beta_{T-t} \log X_T^{\pi^\varepsilon} + \int_{t+\varepsilon}^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^\varepsilon}) - \lambda \log(\pi_{l,X_l^{\pi^\varepsilon}}^*(u, c))) \pi_{l,X_l^{\pi^\varepsilon}}^*(u, c) dc du dl \right] \\ &= \mathbb{E}_{t,x} \left[\mathbb{E}_{t+\varepsilon, X_{t+\varepsilon}^v} [\beta_{T-t} \log X_T^{\pi^*}] \right] \\ &\quad + \mathbb{E}_{t,x} \left[\mathbb{E}_{t+\varepsilon, X_{t+\varepsilon}^v} \left[\int_{t+\varepsilon}^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_{l,X_l^{\pi^*}}^*(u, c))) \pi_{l,X_l^{\pi^*}}^*(u, c) dc du dl \right] \right] \\ &= \mathbb{E}_{t,x} \left[\mathbb{E}_{t+\varepsilon, X_{t+\varepsilon}^v} \left[\beta_{T-t} \log X_T^{\pi^*} + \int_{t+\varepsilon}^T \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right] \right] \\ &= \mathbb{E}_{t,x} [f(t + \varepsilon, X_{t+\varepsilon}^v, t)], \end{aligned}$$

where the last step is by (4.16). Therefore,

$$J^{\pi^\varepsilon}(t, x) = \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^v) - \lambda \log(v(u, c))) v(u, c) dc du dl \right]$$

$$+\mathbb{E}_{t,x} [f(t+\varepsilon, X_{t+\varepsilon}^v, t)] .$$

Since

$$\mathbb{E}_{t,x} [V(t+\varepsilon, X_{t+\varepsilon}^v)] = \mathbb{E}_{t,x} [f(t+\varepsilon, X_{t+\varepsilon}^v, t+\varepsilon)] ,$$

we have that

$$\begin{aligned} & J^{\pi^\varepsilon}(t, x) - \mathbb{E}_{t,x} [V(t+\varepsilon, X_{t+\varepsilon}^v)] \\ &= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} \beta_{l-t} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^v) - \lambda \log(v(u, c))) v(u, c) dc du dl \right] \\ & \quad + \mathbb{E}_{t,x} [f(t+\varepsilon, X_{t+\varepsilon}^v, t) - f(t+\varepsilon, X_{t+\varepsilon}^v, t+\varepsilon)] \\ &= \varepsilon \left(\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(v(u, c))) v(u, c) dc du - \frac{\partial f}{\partial s}(t, x, t) \right) + o(\varepsilon). \end{aligned}$$

On the other hand, by Itô's formula,

$$\begin{aligned} \mathbb{E}_{t,x} [V(t+\varepsilon, X_{t+\varepsilon}^v)] - V(t, x) &= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} \mathcal{L}^v V(l, X_l^v) dl \right] \\ &= \varepsilon \mathcal{L}^v V(t, x) + o(\varepsilon). \end{aligned}$$

Combining the above two equations yields

$$\begin{aligned} & J^{\pi^\varepsilon}(t, x) - V(t, x) \\ &= J^{\pi^\varepsilon}(t, x) - \mathbb{E}_{t,x} [V(t+\varepsilon, X_{t+\varepsilon}^v)] + \varepsilon \mathcal{L}^v V(t, x) + o(\varepsilon) \\ &= \varepsilon \left(\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(v(u, c))) v(u, c) dc du - \frac{\partial f}{\partial s}(t, x, t) + \mathcal{L}^v V(t, x) \right) + o(\varepsilon). \end{aligned} \tag{A.36}$$

We conclude from (4.12) that

$$\limsup_{\varepsilon \downarrow 0} \frac{J^{\pi^\varepsilon}(t, x) - J^{\pi^*}(t, x)}{\varepsilon} = \limsup_{\varepsilon \downarrow 0} \frac{J^{\pi^\varepsilon}(t, x) - V(t, x)}{\varepsilon} \leq 0,$$

which verifies that π^* is an equilibrium policy.

A.3.3 Proof of Lemma 4.4.0.1

It is seen that problem (4.17) is equivalent to the auxiliary problem that

$$\max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(c) - \lambda \log(\pi(u, c)) + \left((\mu - r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) + \lambda K \right) \pi(u, c) dc du, \tag{A.37}$$

where we define the constant

$$\begin{aligned} K &:= 1 - \ln \int_{\mathbb{R}} \int_{\mathbb{R}_+} \frac{c}{\lambda} \exp \left\{ -\frac{A(t)}{\lambda} c \right\} \exp \left\{ \frac{(\mu-r)}{\lambda} u - \frac{\sigma^2 A(t)}{2\lambda} u^2 \right\} dc du \\ &= 1 - \frac{A(t)(\mu-r)^2}{2\lambda\sigma^2} - \ln \left(\frac{\Gamma(\frac{\lambda+1}{\lambda})}{(A(t)/\lambda)^{(\lambda+1)/\lambda}} \sqrt{\frac{2\pi\lambda}{\sigma^2 A(t)}} \right). \end{aligned}$$

Since $\int_{\mathbb{R}} \int_{\mathbb{R}_+} \pi(u, c) dc du = 1$, it is clear that problem (4.17) and problem (A.37) are equivalent. Interchanging the integrals and the maximization yields

$$\begin{aligned} & \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(c) - \lambda \log(\pi(u, c)) + \left((\mu-r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) + \lambda K \right) \pi(u, c) dc du \\ & \leq \int_{\mathbb{R}} \int_{\mathbb{R}_+} \max_{\pi > 0} \left\{ \left(\log(c) - \lambda \log(\pi) + \left((\mu-r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) + \lambda K \right) \pi \right\} dc du. \end{aligned}$$

It is seen that, for fixed u, c , the function

$$\pi \mapsto \left(\log(c) - \lambda \log(\pi) + \left((\mu-r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) + \lambda K \right) \pi, \quad \text{for } \pi > 0,$$

has a unique maximum attained at

$$\begin{aligned} \pi^*(u, c) &= \exp \left\{ \frac{\log(c) + \left((\mu-r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) + \lambda K - \lambda}{\lambda} \right\} \\ &= \frac{c^{\frac{1}{\lambda}} \exp \left\{ -\frac{A(t)}{\lambda} c \right\} \cdot \exp \left\{ \frac{(\mu-r)A(t)}{\lambda} u - \frac{\sigma^2 A(t)}{2\lambda} u^2 \right\}}{\exp \{1 - K\}} \\ &= \frac{c^{\frac{1}{\lambda}} \exp \left\{ -\frac{A(t)}{\lambda} c \right\} \cdot \exp \left\{ -\frac{\sigma^2 A(t)}{2\lambda} \left(u - \frac{\mu-r}{\sigma^2} \right)^2 \right\}}{\exp \{1 - K\} \exp \left\{ -\frac{A(t)(\mu-r)^2}{2\lambda\sigma^2} \right\}}. \end{aligned}$$

From the definition of K , that is,

$$K = 1 - \ln \int_{\mathbb{R}} \int_{\mathbb{R}_+} \frac{c}{\lambda} \exp \left\{ -\frac{A(t)}{\lambda} c \right\} \exp \left\{ \frac{(\mu-r)}{\lambda} u - \frac{\sigma^2 A(t)}{2\lambda} u^2 \right\} dc du,$$

it is straightforward to verify that

$$\exp \{1 - K\} \exp \left\{ -\frac{A(t)(\mu-r)^2}{2\lambda\sigma^2} \right\}$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}_+} c^{\frac{1}{\lambda}} \exp \left\{ -\frac{A(t)}{\lambda} c \right\} \cdot \exp \left\{ -\frac{\sigma^2 A(t)}{2\lambda} \left(u - \frac{\mu - r}{\sigma^2} \right)^2 \right\} dc du,$$

which implies that $\pi^*(u, c)$ is a valid probability density. Moreover, from the separation form of the two variables u and c in $\pi^*(u, c)$, we conclude that

$$\pi^*(u, c) = \pi^{*,U}(u) \cdot \pi^{*,C}(c),$$

where $\pi^{*,U}(u) \sim \text{Normal} \left(\frac{\mu - r}{\sigma^2}, \frac{\lambda}{\sigma^2 A(t)} \right)$ and $\pi^{*,C}(c) \sim \text{Gamma} \left(\frac{\lambda + 1}{\lambda}, \frac{A(t)}{\lambda} \right)$.

A.3.4 Proof of Theorem 4.4.2

Substituting the infinitesimal generator (4.11) into (4.12) yields

$$\begin{aligned} 0 &= \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^\pi V(t, x) - \frac{\partial f}{\partial s}(t, x, t) \right\} \\ &= \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cx) - \lambda \log(\pi(u, c)) + ((\mu - r)u - c)xV_x + \frac{\sigma^2 u^2 x^2}{2} V_{xx} \right) \pi(u, c) dc du \\ &\quad + V_t + rxV_x - \frac{\partial f}{\partial s}(t, x, t) \end{aligned} \tag{A.38}$$

For ease of notation, we omit the arguments (t, x) of V_t , V_x , and V_{xx} .

We consider the following ansatz for the value function

$$V(t, x) = A(t) \log(x) + B(t), \tag{A.39}$$

where $A(\cdot)$ and $B(\cdot)$ are two functions of t to be determined with the boundary condition that $A(T) = 1$ and $B(T) = 0$. Then the EHJB equation (A.38) becomes

$$\begin{aligned} 0 &= \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(c) - \lambda \log(\pi(u, c)) + \left((\mu - r)u - c - \frac{\sigma^2 u^2}{2} \right) A(t) \right) \pi(u, c) dc du \\ &\quad + (A'(t) + 1) \log(x) + B'(t) + rA(t) - \frac{\partial f}{\partial s}(t, x, t). \end{aligned} \tag{A.40}$$

By Lemma 4.4.0.1, we have that the equilibrium investment-consumption policy is given by

$$\pi^*(u, c) = \pi^{*,U}(u) \cdot \pi^{*,C}(c),$$

where

$$\pi^{*,U}(u) \sim \text{Normal}\left(\frac{\mu-r}{\sigma^2}, \frac{\lambda}{\sigma^2 A(t)}\right) \quad \text{and} \quad \pi^{*,C}(c) \sim \text{Gamma}\left(\frac{\lambda+1}{\lambda}, \frac{A(t)}{\lambda}\right). \quad (\text{A.41})$$

To avoid ambiguity, we note that the probability density function of a random variable X that follows $\text{Gamma}(\alpha, \beta)$ is given by $\pi_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$ for $x > 0$. Furthermore, we have $\mathbb{E}[X] = \frac{\alpha}{\beta}$, $\mathbb{E}[\log X] = \phi(\alpha) - \log \beta$, and the entropy of X is given by

$$\mathbb{E}[-\log \pi_X(X)] = - \int_{\mathbb{R}_+} \pi_X(x) \log \pi_X(x) dx = \alpha - \log(\beta) + \log(\Gamma(\alpha)) + (1-\alpha)\phi(\alpha),$$

where $\phi(\alpha) := \frac{d}{d\alpha} \log \Gamma(\alpha)$ is the digamma function.

It remains to solve for the value function $V(t, x)$ and the auxiliary value function $f(t, x, s)$. Substituting the distribution of π^* back into (A.40) and with some simplifications, we obtain

$$\begin{aligned} 0 &= \frac{\lambda}{2} \log\left(\frac{2\pi\lambda}{\sigma^2 A(t)}\right) - (\lambda+1) \log\left(\frac{A(t)}{\lambda}\right) + \lambda \log\left(\Gamma\left(\frac{\lambda+1}{\lambda}\right)\right) + \frac{(\mu-r)^2}{2\sigma^2} A(t) \\ &\quad + (A'(t) + 1) \log(x) + B'(t) + rA(t) - \frac{\partial f}{\partial s}(t, x, t). \end{aligned} \quad (\text{A.42})$$

Next we turn to deriving an expression for $f(t, x, s)$. By (A.41), the portfolio value process under the equilibrium investment policy $\pi^{*,U}(u)$ and the equilibrium consumption policy $\pi^{*,C}(c)$ is given by

$$\begin{aligned} dX_t^{\pi^*} &= \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu-r)u - c) \pi^*(u, c) dc du \right) X_t^{\pi^*} dt + X_t^{\pi^*} \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi^{*,U}(u) du} dW_t \\ &= \left(r + \int_{\mathbb{R}} (\mu-r)u \pi_t^{*,U}(u) du - \int_{\mathbb{R}_+} c \pi_t^{*,C}(c) dc \right) X_t^{\pi^*} dt + X_t^{\pi^*} \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi_t^{*,U}(u) du} dW_t \\ &= \left(r + \frac{(\mu-r)^2}{\sigma^2} - \frac{\lambda+1}{A(t)} \right) X_t^{\pi^*} dt + X_t^{\pi^*} \sqrt{\frac{(\mu-r)^2}{\sigma^2} + \frac{\lambda}{A(t)}} dW_t. \end{aligned}$$

It follows that

$$d \log X_t^{\pi^*} = \left(r + \frac{(\mu-r)^2}{\sigma^2} - \frac{\lambda+1}{A(t)} \right) dt + \sqrt{\frac{(\mu-r)^2}{\sigma^2} + \frac{\lambda}{A(t)}} dW_t$$

$$\begin{aligned}
& -\frac{1}{2} \left(\frac{(\mu - r)^2}{\sigma^2} + \frac{\lambda}{A(t)} \right) dt \\
& = \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(t)} \right) dt + \sqrt{\frac{(\mu - r)^2}{\sigma^2} + \frac{\lambda}{A(t)}} dW_t.
\end{aligned}$$

Thus, for $s \geq t$,

$$\mathbb{E}_{t,x}[\log X_s^{\pi^*}] = \log(x) + \int_t^s \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl. \quad (\text{A.43})$$

By (4.16), (A.41) and the entropy for the normal and gamma distributions, we have that

$$\begin{aligned}
f(t, x, s) &= \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^{\pi^*} + \int_t^T \beta_{l-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cX_l^{\pi^*}) - \lambda \log(\pi_l^*(u, c))) \pi_l^*(u, c) dc du dl \right] \\
&= \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^{\pi^*} + \int_t^T \beta_{l-s} \int_{\mathbb{R}_+} \log(cX_l^{\pi^*}) \pi_l^{*,C}(c) dc dl \right] \\
&\quad - \lambda \mathbb{E}_{t,x} \left[\int_t^T \beta_{l-s} \left(\int_{\mathbb{R}} \pi_l^{*,U}(u) \log \pi_l^{*,U}(u) du + \int_{\mathbb{R}_+} \pi_l^{*,C}(c) \log \pi_l^{*,C}(c) dc \right) dl \right] \\
&= \mathbb{E}_{t,x} \left[\beta_{T-s} \log X_T^{\pi^*} + \int_t^T \beta_{l-s} \log X_l^{\pi^*} dl \right] \\
&\quad + \int_t^T \beta_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3}{2}\lambda + 1 + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) \right) dl.
\end{aligned}$$

Substituting (A.43) into the above expression yields

$$\begin{aligned}
f(t, x, s) &= \left(\beta_{T-s} + \int_t^T \beta_{l-s} dl \right) \log(x) + \beta_{T-s} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl \\
&\quad + \int_t^T \beta_{l-s} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
&\quad + \int_t^T \beta_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) \right) dl.
\end{aligned}$$

It follows that

$$-\frac{\partial f}{\partial s}(t, x, t)$$

$$\begin{aligned}
&= [\beta'_{T-t} + (\beta_{T-t} - 1)] \log(x) + \beta'_{T-t} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl \\
&\quad + \int_t^T \beta'_{l-t} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
&\quad + \int_t^T \beta'_{l-t} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) \right) dl.
\end{aligned}$$

Substituting this into (A.42) yields

$$\begin{aligned}
0 &= \log(x) [A'(t) + \beta'_{T-t} + \beta_{T-t}] \\
&\quad + B'(t) + \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(t) - (\lambda + 1) \log \left(\frac{A(t)}{\lambda} \right) + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(t)} \right) \\
&\quad + \beta'_{T-t} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl + \int_t^T \beta'_{l-t} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
&\quad + \int_t^T \beta'_{l-t} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) \right) dl.
\end{aligned}$$

It follows that

$$A'(t) + \beta'_{T-t} + \beta_{T-t} = 0,$$

and

$$\begin{aligned}
0 &= B'(t) + \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(t) - (\lambda + 1) \log \left(\frac{A(t)}{\lambda} \right) + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(t)} \right) \\
&\quad + \beta'_{T-t} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl + \int_t^T \beta'_{l-t} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
&\quad + \int_t^T \beta'_{l-t} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log \left(\Gamma \left(\frac{\lambda + 1}{\lambda} \right) \right) \right) dl.
\end{aligned}$$

This completes the proof by the boundary conditions that $A(T) = 1$ and $B(T) = 0$.

A.3.5 Proof of Corollary 4.4.1

We first show the convergence of policies. From (4.18), the Gaussian investment policy has a constant mean $\frac{\mu - r}{\sigma^2}$ and a variance

$$\frac{\lambda}{\sigma^2 A(t)} \rightarrow 0,$$

as $\lambda \rightarrow 0$. So, $\pi_{t,x}^*(u)$ converges in distribution to $\frac{\mu-r}{\sigma^2}$, which is the equilibrium investment policy in the classical, non-exploratory problem in Theorem 4.2.1.

Also from (4.18), the Gamma consumption policy has mean $\frac{\lambda+1}{A(t)}$ and variance $\frac{\lambda(\lambda+1)}{A(t)}$. As $\lambda \rightarrow 0$,

$$\begin{aligned}\frac{\lambda+1}{A(t)} &\rightarrow \frac{1}{A(t)} \\ \frac{\lambda(\lambda+1)}{A(t)} &\rightarrow 0.\end{aligned}$$

So, $\pi_{t,x}^*(c)$ converges in distribution to $\frac{1}{A(t)}$, which is the equilibrium consumption policy in the classical, non-exploratory problem in Theorem 4.2.1.

We next show the convergence of the corresponding value function. Recall that under both exploratory and non-exploratory setups, the value function takes the form $V(t, x) = A(t) \log(x) + B(t)$. It is clear that $A(t)$ is the same under both setups and it does not depend on λ . From Theorem 4.2.1 and Theorem 4.4.2, we see that

$$\begin{aligned}b(s) = & b_0(s) - \lambda \log A(s) + (\lambda + 1) \log \lambda + \lambda \log(\Gamma(1 + \frac{1}{\lambda})) + \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(s)} \right) \\ & + \int_s^T \beta'_{t-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) + \lambda \log \left(\frac{1}{A(l)} \right) \right) dl \\ & + (\beta_{T-s} - 1) [(\lambda + 1) \log \lambda + \frac{3}{2} \lambda + 1 + \lambda \log \Gamma(1 + \frac{1}{\lambda})] + 1.\end{aligned}$$

The following lemma is sufficient to show that $b(s) \rightarrow b_0(s)$, as $\lambda \rightarrow 0^+$.

Lemma A.3.5.1

$$\lim_{x \rightarrow 0^+} \left(\log(x) + x \log \left(\Gamma(1 + \frac{1}{x}) \right) \right) = -1.$$

We herein provide a sketch proof of this Lemma. For large y , Stirling's Approximation (Montgomery & Vaughan, 2007) yields

$$\log(\Gamma(y)) = y \log(y) - y + \frac{1}{2} \log(2\pi y) + O(1/|y|).$$

Replacing y with $1 + \frac{1}{x}$,

$$\log \left(\Gamma \left(1 + \frac{1}{x} \right) \right) = \left(1 + \frac{1}{x} \right) \log \left(1 + \frac{1}{x} \right) - \left(1 + \frac{1}{x} \right) + \frac{1}{2} \log \left(2\pi \left(1 + \frac{1}{x} \right) \right) + O \left(\left| \frac{x}{x+1} \right| \right).$$

For small x , we also have

$$\log\left(1 + \frac{1}{x}\right) = \log\left(\frac{1}{x}\right) + \log(1 + x) = \log\left(\frac{1}{x}\right) + x + O(x^2).$$

So,

$$\begin{aligned} \log\left(\Gamma\left(1 + \frac{1}{x}\right)\right) &= \left(1 + \frac{1}{x}\right) \left(\log\left(\frac{1}{x}\right) + x + O(x^2)\right) - \left(1 + \frac{1}{x}\right) + \frac{1}{2} \log(2\pi) \\ &\quad + \frac{1}{2} \left(\log\left(\frac{1}{x}\right) + x + O(x^2)\right) + O\left(\left|\frac{x}{x+1}\right|\right) \\ &= \left(1 + \frac{1}{x}\right) \log\left(\frac{1}{x}\right) + (x+1) + O(x) - \left(1 + \frac{1}{x}\right) + \frac{1}{2} \log\left(\frac{1}{x}\right) \\ &\quad + \frac{x}{2} + O(x^2) + \frac{1}{2} \log(2\pi) + O\left(\left|\frac{x}{x+1}\right|\right) \\ &= \left(\frac{3}{2} + \frac{1}{x}\right) \log\left(\frac{1}{x}\right) + \frac{3x}{2} - \frac{1}{x} + \frac{1}{2} \log(2\pi) + O(x) + O(x^2) + O\left(\left|\frac{x}{x+1}\right|\right). \end{aligned}$$

This implies,

$$\begin{aligned} &\log(x) + x \log\left(\Gamma\left(1 + \frac{1}{x}\right)\right) \\ &= -\frac{3}{2}x \log(x) + \frac{3}{2}x^2 + \frac{x}{2} \log(2\pi) - 1 + O(x^2) + O(x^3) + O\left(\left|\frac{x^2}{x+1}\right|\right) \\ &= -\frac{3}{2}x(x-1) + \frac{3}{2}x^2 + \frac{x}{2} \log(2\pi) - 1 + O(x^2) + O(x^3) + O\left(\left|\frac{x^2}{x+1}\right|\right), \end{aligned}$$

where the last equation is given by $\log x = (x-1) + O(x^2)$. So,

$$\lim_{x \rightarrow 0^+} \left(\log(x) + x \log\left(\Gamma\left(1 + \frac{1}{x}\right)\right) \right) = -1.$$

A.3.6 Proof of Theorem 4.4.3

From Theorem 4.4.1, we know that $V^{(0)}$ and $f^{(0)}$ satisfy the following conditions

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi(u, c))) \pi^{(0)}(u, c) dc du + \mathcal{L}^{\pi^{(0)}} V^{(0)}(t, x) - \frac{\partial f^{(0)}}{\partial s}(t, x, t) = 0,$$

and

$$\beta_{t-s} \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(cx) - \lambda \log(\pi^{(0)}(u, c))) \pi^{(0)}(u, c) dc du + \mathcal{L}^{\pi^{(0)}} f^{(0)}(t, x, s) = 0,$$

with boundary condition $V^{(0)}(T, x) = \log x$ and $f^{(0)}(T, x, s) = \beta_{T-s} \log x$. By a similar derivation as in the proof of Theorem 4.4.2, we can show that

$$V^{(0)}(t, x) = A(t) \log(x) + B^{(0)}(t), \quad (\text{A.44})$$

where $A(t)$ is given in (4.19) and $B^{(0)}(t)$ is a deterministic function of t .

By (4.23) and (A.44),

$$\begin{aligned} \pi^{(1)} &= \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(c) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + \mathcal{L}^{\pi} V^{(0)}(t, x) \right\} \\ &= \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} (\log(c) - \lambda \log(\pi(u, c))) \pi(u, c) dc du + A'(t) \log(x) + B^{(0)}(t)' \right. \\ &\quad \left. + A(t) \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left((\mu - r)u - c - \frac{\sigma^2 u^2}{2} \right) \pi(u, c) dc du \right) \right\} \\ &= \arg \max_{\pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R}_+)} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left[\log(c) - \lambda \log(\pi(u, c)) \right. \right. \\ &\quad \left. \left. + A(t) \left((\mu - r)u - c - \frac{\sigma^2 u^2}{2} \right) \right] \pi(u, c) dc du \right\}. \end{aligned}$$

By Lemma 4.4.0.1, we conclude that

$$\pi^{(1)} = \pi^*,$$

where π^* is the equilibrium investment-consumption policy in Theorem 4.4.2.

A.3.7 Proof of Theorem 4.5.1

By (4.26), we have

$$M_t^\theta = A(t) \log(X_t^{\pi^\theta}) + \int_t^T b^\theta(s) ds$$

$$- \int_0^t \left(- \int_{\mathbb{R}_+} \log(cX_s^{\pi^\theta}) \pi_s^{\theta,C}(c) + \lambda \int_{\mathbb{R}} \int_{\mathbb{R}_+} \pi_s^\theta(u, c) \log \pi_s^\theta(u, c) dc du + \frac{\partial f^\theta}{\partial s}(s, X_s^{\pi^\theta}, s) \right) ds. \quad (\text{A.45})$$

It follows from (4.24) that

$$\begin{aligned} & - \int_{\mathbb{R}_+} \log(cX_s^{\pi^\theta}) \pi_s^{\theta,C}(c) + \lambda \int_{\mathbb{R}} \int_{\mathbb{R}_+} \pi_s^\theta(u, c) \log \pi_s^\theta(u, c) dc du \\ = & - \log X_s^{\pi^\theta} - \left(\phi\left(\frac{\lambda+1}{\lambda}\right) - \log\left(\frac{A(s)}{\lambda}\right) \right) \\ & - \lambda \left(\frac{1}{2} \log\left(\frac{2\pi e \lambda}{\sigma^2 A(s)}\right) + \frac{\lambda+1}{\lambda} - \log\left(\frac{A(s)}{\lambda}\right) + \log(\Gamma(\frac{\lambda+1}{\lambda})) + (1 - \frac{\lambda+1}{\lambda}) \phi\left(\frac{\lambda+1}{\lambda}\right) \right) \\ = & - \log X_s^{\pi^\theta} + (\lambda+1) \log\left(\frac{A(s)}{\lambda}\right) - \frac{\lambda}{2} \log\left(\frac{2\pi \lambda}{\sigma^2 A(s)}\right) - \frac{3\lambda+2}{2} - \lambda \log\left(\Gamma\left(\frac{\lambda+1}{\lambda}\right)\right). \end{aligned} \quad (\text{A.46})$$

By (4.28)

$$\begin{aligned} & \frac{\partial f^\theta(s, X_s^{\pi^\theta}, s)}{\partial s} \\ = & - (\beta'_{T-s} + \beta_{T-s} - 1) \log(X_s^{\pi^\theta}) - \beta'_{T-s} \int_s^T \left(r + \frac{(\mu-r)^2}{2\sigma^2} - \frac{3\lambda+2}{2A(l)} \right) dl \\ & - \int_s^T \beta'_{l-s} \int_s^l \left(r + \frac{(\mu-r)^2}{2\sigma^2} - \frac{3\lambda+2}{2A(w)} \right) dw dl \\ & - \int_s^T \beta'_{l-s} \left(\frac{\lambda}{2} \log\left(\frac{2\pi \lambda}{\sigma^2 A(l)}\right) - (\lambda+1) \log\left(\frac{A(l)}{\lambda}\right) + \frac{3\lambda+2}{2} + \lambda \log(\Gamma(\frac{\lambda+1}{\lambda})) \right) dl. \end{aligned} \quad (\text{A.47})$$

Substituting the above two expressions together with (4.27) into (A.45) yields

$$\begin{aligned} M_t^\theta &= A(t) \log(X_t^{\pi^\theta}) + \int_t^T b^\theta(s) ds \\ & - \int_0^t \left\{ (\lambda+1) \log\left(\frac{A(s)}{\lambda}\right) - \frac{\lambda}{2} \log\left(\frac{2\pi \lambda}{\sigma^2 A(s)}\right) - \frac{3\lambda+2}{2} - \lambda \log\left(\Gamma\left(\frac{\lambda+1}{\lambda}\right)\right) \right. \\ & \left. - (\beta'_{T-s} + \beta_{T-s}) \log(X_s^{\pi^\theta}) + \beta'_{T-s} \int_s^T \left(r + \frac{(\mu-r)^2}{2\sigma^2} - \frac{3\lambda+2}{2A(l)} \right) dl \right\} ds \end{aligned}$$

$$\begin{aligned}
& - \int_s^T \beta'_{l-s} \int_s^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
& - \int_s^T \beta'_{l-s} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log(\Gamma(\frac{\lambda + 1}{\lambda})) \right) dl \Big\} ds.
\end{aligned}$$

By (4.25) and $A'(t) + \beta'_{T-t} + \beta_{T-t} = 0$, we have that

$$\begin{aligned}
dM_t^\theta &= A'(t) \log(X_t^{\pi^\theta}) dt + A(t) \left(\frac{1}{X_t^{\pi^\theta}} dX_t^{\pi^\theta} - \frac{1}{2} \frac{1}{(X_t^{\pi^\theta})^2} (dX_t^{\pi^\theta})^2 \right) - b^\theta(t) dt \\
& - \left\{ (\lambda + 1) \log \left(\frac{A(t)}{\lambda} \right) - \frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(t)} \right) - \frac{3\lambda + 2}{2} - \lambda \log \left(\Gamma(\frac{\lambda + 1}{\lambda}) \right) \right. \\
& - (\beta'_{T-t} + \beta_{T-t}) \log(X_t^{\pi^\theta}) + \beta'_{T-t} \int_t^T \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(l)} \right) dl \\
& - \int_t^T \beta'_{l-t} \int_t^l \left(r + \frac{(\mu - r)^2}{2\sigma^2} - \frac{3\lambda + 2}{2A(w)} \right) dw dl \\
& \left. - \int_t^T \beta'_{l-t} \left(\frac{\lambda}{2} \log \left(\frac{2\pi\lambda}{\sigma^2 A(l)} \right) - (\lambda + 1) \log \left(\frac{A(l)}{\lambda} \right) + \frac{3\lambda + 2}{2} + \lambda \log(\Gamma(\frac{\lambda + 1}{\lambda})) \right) dl \right\} dt \\
&= \left[A(t) \left(r + \frac{(\mu_{true} - r)(\mu - r)}{\sigma^2} - \frac{\lambda + 1}{A(t)} \right) - \frac{A(t)}{2} \left(\frac{(\mu - r)^2 \sigma_{true}^2}{\sigma^4} + \frac{\lambda \sigma_{true}^2}{\sigma^2 A(t)} \right) \right. \\
& \quad \left. + \frac{3\lambda + 2}{2} - \left(r + \frac{(\mu - r)^2}{2\sigma^2} \right) A(t) \right] dt \\
& \quad + A(t) \sigma_{true} \sqrt{\frac{(\mu - r)^2}{\sigma^4} + \frac{\lambda}{\sigma^2 A(t)}} dW_t.
\end{aligned}$$

Therefore, M^θ is a martingale if $\theta = (\mu, \sigma) = (\mu_{true}, \sigma_{true})$.

A.3.8 Proof of Theorem 4.5.2

Recall that

$$N_t^\theta := \beta_t V^\theta(t, X_t^{\pi^\theta}) + \int_0^t \beta_s \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_s^{\pi^\theta}) - q(s, X_s^{\pi^\theta}, (u, c); \pi^\theta) \right) \pi_s^\theta(u, c) dc du ds$$

$$- \int_0^t \left(\beta'_s V^\theta(s, X_s^{\pi^\theta}) + \beta_s \frac{\partial f^\theta}{\partial s}(s, X_s^{\pi^\theta}, s) \right) ds.$$

By Itô's formula,

$$\begin{aligned} dN_t^\theta &= \beta_t dV^\theta(t, X_t^{\pi^\theta}) + \beta_t \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_t^{\pi^\theta}) - q(t, X_t^{\pi^\theta}, (u, c); \pi^\theta) \right) \pi_t^\theta(u, c) dc du dt \\ &\quad - \beta_t \frac{\partial f^\theta}{\partial s}(t, x, t) dt \\ &= \beta_t \frac{\partial V^\theta(t, X_t^{\pi^\theta})}{\partial t} dt + \beta_t \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu_{true} - r)u - c) \pi_t^\theta(u, c) dc du \right) X_t^{\pi^\theta} \frac{\partial V^\theta(t, X_t^{\pi^\theta})}{\partial x} dt \\ &\quad + \beta_t \frac{(X_t^{\pi^\theta})^2}{2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \sigma_{true}^2 u^2 \pi_t^\theta(u, c) dc du \frac{\partial^2 V^\theta(t, X_t^{\pi^\theta})}{\partial x^2} dt \\ &\quad + \beta_t \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_t^{\pi^\theta}) - q(s, X_s^{\pi^\theta}, (u, c); \pi^\theta) \right) \pi_t^\theta(u, c) dc du dt - \beta_t \frac{\partial f^\theta}{\partial s}(t, x, t) dt \\ &\quad + \beta_t \frac{\partial V^\theta(t, X_t^{\pi^\theta})}{\partial x} X_t^{\pi^\theta} \sqrt{\int_{\mathbb{R}} \sigma_{true}^2 u^2 \pi_t^{\theta, U}(u) du} dW_t. \end{aligned}$$

After factoring out the coefficient β_t , the drift (dt) term of dN_t^θ is given by

$$\begin{aligned} &\frac{\partial V^\theta(t, X_t^{\pi^\theta})}{\partial t} + \left(r + \int_{\mathbb{R}} \int_{\mathbb{R}_+} ((\mu_{true} - r)u - c) \pi_t^\theta(u, c) dc du \right) X_t^{\pi^\theta} \frac{\partial V^\theta(t, X_t^{\pi^\theta})}{\partial x} \\ &+ \frac{(X_t^{\pi^\theta})^2}{2} \int_{\mathbb{R}} \int_{\mathbb{R}_+} \sigma_{true}^2 u^2 \pi_t^\theta(u, c) dc du \frac{\partial^2 V^\theta(t, X_t^{\pi^\theta})}{\partial x^2} \\ &+ \int_{\mathbb{R}} \int_{\mathbb{R}_+} \left(\log(cX_t^{\pi^\theta}) - q(s, X_s^{\pi^\theta}, (u, c); \pi^\theta) \right) \pi_t^\theta(u, c) dc du dt - \frac{\partial f^\theta}{\partial s}(t, x, t). \end{aligned}$$

When $\theta = (\mu_{true}, \sigma_{true})$, by (4.35), the drift term of dN_t^θ vanishes, and therefore N^θ is a martingale.

A.4 Proofs of Chapter 5

A.4.1 Derivation of (5.26) and (5.27)

Note that for any smooth function $\psi(t, x) \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R})$ and under Assumption 5.2.1, Itô's formula implies

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \left[\mathbb{E}_{t,x} [\psi(t + \Delta t, X_{t+\Delta t})] - \psi(t, x) \right] = \mathcal{L}\psi(t, x).$$

Suppose J_{ne}^π and f^π are smooth with $J_{ne}^\pi \in \mathbb{C}^{1,2}([0, T] \times \mathbb{R})$ and $f^\pi \in \mathbb{C}^{1,2,1}([0, T] \times \mathbb{R} \times [0, t])$. Then, we have

$$\begin{aligned} \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \left[\mathbb{E}_{t,x} [J_{ne}^\pi(t + \Delta t, X_{t+\Delta t})] - J_{ne}^\pi(t, x) \right] &= \mathcal{L}J_{ne}^\pi(t, x), \\ \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \left[\mathbb{E}_{t,x,s} [f^\pi(t + \Delta t, X_{t+\Delta t}, s)] - f^\pi(t, x, s) \right] &= \mathcal{L}f^\pi(t, x, s). \end{aligned}$$

To derive (5.26), we start from the objective function (5.23), for $(t, x) \in [0, T] \times \mathbb{R}$,

$$\begin{aligned} & J_{ne}^\pi(t, x) \\ = & \mathbb{E}_{t,x} \left\{ \int_t^{t+\Delta t} e^{-K \int_t^s \pi_l dl} \beta_{s-t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right. \\ & + \int_{t+\Delta t}^T e^{-K \int_t^s \pi_l dl} \beta_{s-t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \\ & + e^{-K \int_t^T \pi_l dl} \beta_{T-t} g(X_T) \\ & + \int_{t+\Delta t}^T e^{-K \int_{t+\Delta t}^s \pi_l dl} \beta_{s-t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \\ & + e^{-K \int_{t+\Delta t}^T \pi_l dl} \beta_{T-t} g(X_T) \\ & - \int_{t+\Delta t}^T e^{-K \int_{t+\Delta t}^s \pi_l dl} \beta_{s-t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \\ & - e^{-K \int_{t+\Delta t}^T \pi_l dl} \beta_{T-t} g(X_T) \\ & + J_{ne}^\pi(t + \Delta t, X_{t+\Delta t}) \\ & \left. - \mathbb{E}_{t+\Delta t, X_{t+\Delta t}} \left[\int_{t+\Delta t}^T e^{-K \int_{t+\Delta t}^s \pi_l dl} \beta_{s-t-\Delta t} \left(Kg(X_s) \pi_s - \lambda(\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right] \right\} \end{aligned}$$

$$\begin{aligned}
& -\mathbb{E}_{t+\Delta t, X_{t+\Delta t}} \left[e^{-K \int_{t+\Delta t}^T \pi_l dl} \beta_{T-t-\Delta t} g(X_T) \right] \Bigg\} \\
= & \mathbb{E}_{t,x} \left\{ \int_t^{t+\Delta t} e^{-K \int_t^s \pi_l dl} \beta_{s-t} \left(K g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \right. \\
& + \int_{t+\Delta t}^T e^{-K \int_t^s \pi_l dl} \beta_{s-t} \left(K g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \\
& + e^{-K \int_t^T \pi_l dl} \beta_{T-t} g(X_T) \\
& - \int_{t+\Delta t}^T e^{-K \int_{t+\Delta t}^s \pi_l dl} \beta_{s-t} \left(g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s)) \right) ds \\
& - e^{-K \int_{t+\Delta t}^T \pi_l dl} \beta_{T-t} g(X_T) \\
& \left. + f^\pi(t + \Delta t, X_{t+\Delta t}, t) - f^\pi(t + \Delta t, X_{t+\Delta t}, t + \Delta t) + J_{ne}^\pi(t + \Delta t, X_{t+\Delta t}) \right\}.
\end{aligned}$$

Dividing by Δt and taking $\Delta t \rightarrow 0$, the above equation becomes

$$\begin{aligned}
& [K g(x) \pi_t - \lambda (\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] \\
& - K \pi_t \mathbb{E}_{t,x} \left[\int_t^T K e^{-K \int_t^s \pi_l dl} \beta_{s-t} (g(X_s) \pi_s - \lambda (\pi_s \log \pi_s + (1 - \pi_s) \log(1 - \pi_s))) ds \right. \\
& \left. + e^{-K \int_t^T \pi_l dl} \beta_{T-t} g(X_T) \right] \\
& + \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left(\mathbb{E}_{t,x} [f^\pi(t + \Delta t, X_{t+\Delta t}, t) - f^\pi(t + \Delta t, X_{t+\Delta t}, t + \Delta t)] \right) \\
& + \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \left[\mathbb{E}_{t,x} [J_{ne}^\pi(t + \Delta t, X_{t+\Delta t})] - J_{ne}^\pi(t, x) \right] \\
= & K g(x) \pi_t - \lambda (\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t)) - K \pi_t J_{ne}^\pi - \frac{\partial f^\pi}{\partial s}(t, x, t) + \mathcal{L} J_{ne}^\pi = 0.
\end{aligned}$$

Taking the maximum over π yields (5.26), where J_{ne}^π is replaced by $V(t, x) = \max_{\pi \in [0,1]} J_{ne}^\pi(t, x)$ and f^π is replaced by $f(t, x, s) = f^{\pi^*}(t, x, s)$ with $\pi^* = \arg \max_{\pi \in [0,1]} J_{ne}^\pi(t, x)$.

To derive (5.27), consider a given policy π . From the auxiliary objective function (5.25), for $(t, x, s) \in [0, T] \times \mathbb{R} \times [0, t]$,

$$f^\pi(t, x, s)$$

$$\begin{aligned}
&= \mathbb{E}_{t,x,s} \left\{ \int_t^{t+\Delta t} e^{-K \int_t^v \pi_l dl} \beta_{v-s} \left[Kg(X_v) \pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right] dv \right. \\
&\quad + \int_{t+\Delta}^T e^{-K \int_t^v \pi_l dl} \beta_{v-s} \left[Kg(X_v) \pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right] dv \\
&\quad + e^{-K \int_t^T \pi_l dl} \beta_{T-s} g(X_T^\pi) \\
&\quad + f^\pi(t + \Delta t, X_{t+\Delta t}, s) \\
&\quad - \mathbb{E}_{t+\Delta t, X_{t+\Delta t}, s} \left[\int_{t+\Delta}^T e^{-K \int_{t+\Delta}^v \pi_l dl} \beta_{v-s} \left[Kg(X_v) \pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v)) \right] dv \right] \\
&\quad \left. - \mathbb{E}_{t+\Delta t, X_{t+\Delta t}, s} \left[e^{-K \int_{t+\Delta}^T \pi_l dl} \beta_{T-s} g(X_T^\pi) \right] \right\}.
\end{aligned}$$

Dividing by Δt , taking $\Delta t \rightarrow 0$ and replacing π with π^* , the above equation becomes

$$\begin{aligned}
&\beta_{t-s} \left[Kg(x) \pi_t^* - \lambda(\pi_t^* \log \pi_t^* + (1 - \pi_t^*) \log(1 - \pi_t^*)) \right] \\
&\quad - K \pi_t^* \mathbb{E}_{t,x,s} \left[\int_t^T e^{-K \int_t^v \pi_l^* dl} \beta_{v-s} [Kg(X_v) \pi_v^* \right. \\
&\quad \left. - \lambda(\pi_v^* \log \pi_v^* + (1 - \pi_v^*) \log(1 - \pi_v^*))] dv + e^{-K \int_t^T \pi_l^* dl} \beta_{T-s} g(X_T) \right] \\
&\quad + \lim_{\Delta t \downarrow 0} \mathbb{E}_{t,x,s} [f(t + \Delta t, X_{t+\Delta t}, s)] - f(t, x, s) \\
&= \beta_{T-s} (Kg(x) \pi_t^* - \lambda(\pi_t^* \log \pi_t^* + (1 - \pi_t^*) \log(1 - \pi_t^*))) - K \pi_t^* f(t, x, s) \\
&\quad + \mathcal{L}f(t, x, s) = 0,
\end{aligned}$$

which yields (5.27).

A.4.2 Proof of Theorem 5.3.1

We know from Appendix A.4.1 that $V(t, x) := J_{ne}^{\pi^*}(t, x)$ and $f(t, x, s) := f^{\pi^*}(t, x, s)$ satisfy (5.26) and (5.27), where $\pi^* := \pi_{ne}^*$, given by (5.32), is the maximizer of the supremum part of (5.26). So, what remains is to show that π^* is an equilibrium policy using Definition 5.3.1.

Let $\pi_{ne}^{\varepsilon,v}$ be a perturbed stopping policy of π_{ne}^* , similar to the one in Definition 5.3.1. By (5.23), for $(t, x) \in [0, T] \times \mathbb{R}$,

$$J_{ne}^{\pi_{ne}^{\varepsilon,v}}(t, x)$$

$$\begin{aligned}
&= \mathbb{E}_{t,x} \left[\int_t^T e^{-K \int_t^s \pi_{ne,l}^{\varepsilon,v} dl} \beta_{s-t} \left(Kg(X_s) \pi_{ne,s}^{\varepsilon,v} - \lambda(\pi_{ne,s}^{\varepsilon,v} \log \pi_{ne,s}^{\varepsilon,v} + (1 - \pi_{ne,s}^{\varepsilon,v}) \log(1 - \pi_{ne,s}^{\varepsilon,v})) \right) ds \right. \\
&\quad \left. + e^{-K \int_t^T \pi_{ne,l}^{\varepsilon,v} dl} \beta_{T-t} g(X_T) \right] \\
&= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K \int_t^s \pi_{ne,l}^{\varepsilon,v} dl} \beta_{s-t} \left(Kg(X_s) \pi_{ne,s}^{\varepsilon,v} - \lambda(\pi_{ne,s}^{\varepsilon,v} \log \pi_{ne,s}^{\varepsilon,v} + (1 - \pi_{ne,s}^{\varepsilon,v}) \log(1 - \pi_{ne,s}^{\varepsilon,v})) \right) ds \right] \\
&\quad + \mathbb{E}_{t,x} \left[\int_{t+\varepsilon}^T e^{-K \int_t^s \pi_{ne,l}^{\varepsilon,v} dl} \beta_{s-t} \left(Kg(X_s) \pi_{ne,s}^{\varepsilon,v} - \lambda(\pi_{ne,s}^{\varepsilon,v} \log \pi_{ne,s}^{\varepsilon,v} + (1 - \pi_{ne,s}^{\varepsilon,v}) \log(1 - \pi_{ne,s}^{\varepsilon,v})) \right) ds \right. \\
&\quad \left. + e^{-K \int_t^T \pi_{ne,l}^{\varepsilon,v} dl} \beta_{T-t} g(X_T) \right] \\
&= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K(s-t)v} \beta_{s-t} \left(Kg(X_s) v - \lambda(v \log v + (1 - v) \log(1 - v)) \right) ds \right] \\
&\quad + e^{-K\varepsilon v} \mathbb{E}_{t,x} \left[\int_{t+\varepsilon}^T e^{-K \int_{t+\varepsilon}^s \pi_{ne,l}^{\varepsilon,v} dl} \beta_{s-t} \left(Kg(X_s) \pi_{ne,s}^{\varepsilon,v} \right. \right. \\
&\quad \left. \left. - \lambda(\pi_{ne,s}^{\varepsilon,v} \log \pi_{ne,s}^{\varepsilon,v} + (1 - \pi_{ne,s}^{\varepsilon,v}) \log(1 - \pi_{ne,s}^{\varepsilon,v})) \right) ds + e^{-K \int_{t+\varepsilon}^T \pi_{ne,l}^{\varepsilon,v} dl} \beta_{T-t} g(X_T) \right] \\
&= e^{-K\varepsilon v} \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K(s-t-\varepsilon)v} \beta_{s-t} \left(Kg(X_s) v - \lambda(v \log v + (1 - v) \log(1 - v)) \right) \right. \\
&\quad \left. + f(t + \varepsilon, X_{t+\varepsilon}, t) \right].
\end{aligned}$$

Since

$$\mathbb{E}_{t,x}[V_{ne}(t + \varepsilon, X_{t+\varepsilon})] = \mathbb{E}_{t,x}[f(t + \varepsilon, X_{t+\varepsilon}, t + \varepsilon)],$$

we have

$$\begin{aligned}
&J_{ne}^{\pi_{ne}^{\varepsilon,v}}(t, x) - e^{-K\varepsilon v} \mathbb{E}_{t,x}[V_{ne}(t + \varepsilon, X_{t+\varepsilon})] \\
&= e^{-K\varepsilon v} \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K(s-t-\varepsilon)v} \beta_{s-t} \left(Kg(X_s) v - \lambda(v \log v + (1 - v) \log(1 - v)) \right) \right. \\
&\quad \left. + f(t + \varepsilon, X_{t+\varepsilon}, t) - f(t + \varepsilon, X_{t+\varepsilon}, t + \varepsilon) \right]
\end{aligned}$$

$$= e^{-K\varepsilon v} \cdot \varepsilon \left(Kg(X_s)v - \lambda(v \log v + (1-v) \log(1-v)) - \frac{\partial f}{\partial s}(t, x, t) \right) + o(\varepsilon).$$

On the other hand, by Itô's formula,

$$\begin{aligned} & \mathbb{E}_{t,x}[e^{-K\varepsilon v} V_{ne}(t+\varepsilon, X_{t+\varepsilon})] - V_{ne}(t, x) \\ &= \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K(s-t)v} \mathcal{L}V_{ne}(s, X_s) ds + \int_t^{t+\varepsilon} -Kve^{-K(s-t)v} ds \right] \\ &= e^{-K\varepsilon v} \mathbb{E}_{t,x} \left[\int_t^{t+\varepsilon} e^{-K(s-t-\varepsilon)v} \mathcal{L}V_{ne}(s, X_s) ds \right] + o(\varepsilon) \\ &= e^{-K\varepsilon v} \cdot \varepsilon \mathcal{L}V_{ne}(t, x) + o(\varepsilon). \end{aligned}$$

Therefore,

$$\begin{aligned} & J_{ne}^{\pi^{\varepsilon,v}}(t, x) - V_{ne}(t, x) \\ &= e^{-K\varepsilon v} \cdot \varepsilon \left(Kg(X_s)v - \lambda(v \log v + (1-v) \log(1-v)) - \frac{\partial f}{\partial s}(t, x, t) + \mathcal{L}V_{ne}(t, x) \right) + o(\varepsilon), \end{aligned}$$

which, by (5.26), implies that

$$\limsup_{\varepsilon \downarrow 0} \frac{J_{ne}^{\pi^{\varepsilon,v}}(t, x) - J_{ne}^{\pi^*}(t, x)}{\varepsilon} = \limsup_{\varepsilon \downarrow 0} \frac{J_{ne}^{\pi^{\varepsilon,v}}(t, x) - V_{ne}(t, x)}{\varepsilon} \leq 0.$$

A.4.3 Proof of Proposition 5.3.1

Recall that the stopping regions are respectively $D_{exp} = \{(t, x) \in [0, T] \times \mathbb{R} : V_{exp}(t, x) = g(x)\}$ for exponential discounting and $D_{ne} = \{(t, x) \in [0, T] \times \mathbb{R} : V_{ne}(t, x) = g(x)\}$ for non-exponential discounting. If $V_{ne} \geq V_{exp}$, then for any $(t, x) \in D_{ne}$, $V_{exp}(t, x) \leq g(x)$ and so $(t, x) \in D_{exp}$. Meanwhile, there exists $(\tilde{t}, \tilde{x}) \in D_{exp}$ such that $V_{ne}(\tilde{t}, \tilde{x}) \geq g(\tilde{t}, \tilde{x})$ and so $(t, x) \notin D_{ne}$ when the inequality is strict. This implies $D_{ne} \subseteq D_{exp}$. What remains is to show that $V_{ne} \geq V_{exp}$.

With the penalty approximation as described in Section 5.2 (or originally in (Dai et al., 2024)), consider a given penalty parameter $K > 0$. The penalty approximated HJB equations are

$$\text{exponential discounting: } \mathcal{L}V_{exp}^K - \rho V_{exp}^K + K(g - V_{exp}^K)^+ = 0 \quad (\text{A.48})$$

$$\text{non-exponential discounting: } \mathcal{L}V_{ne}^K - \frac{\partial f^K}{\partial s}(t, x, t) + K(g - V_{ne}^K)^+ = 0, \quad (\text{A.49})$$

with terminal conditions $V_{exp}^K(T, x) = g(x)$ and $V_{ne}^K(T, x) = g(x)$. Here, V_{exp}^K, V_{ne}^K with superscript K denote the value functions that solve the penalty approximated HJB equations. Furthermore, f^K is the corresponding auxiliary value function of V_{ne}^K .

For some $(t, x) \in [0, T] \times \mathbb{R}_+$, recalling that $\beta_0^{\text{nonexp}} = 1$, we have

$$V_{ne}^K(t, x) = \sup_{\tau \in [t, T]} \mathbb{E}_{t, x} \left[\frac{\beta_{\tau}^{\text{nonexp}}}{\beta_t^{\text{nonexp}}} g(\tau, X_{\tau}) \right], \quad (\text{A.50})$$

$$f^K(t, x, s) = \sup_{\tau \in [t, T]} \mathbb{E}_{t, x} \left[\frac{\beta_{\tau}^{\text{nonexp}}}{\beta_s^{\text{nonexp}}} g(\tau, X_{\tau}) \right]. \quad (\text{A.51})$$

Then,

$$\begin{aligned} \frac{\partial f^K}{\partial s}(t, x, s) &= \frac{\partial}{\partial s} \sup_{\tau \in [t, T]} \mathbb{E}_{t, x} \left[\frac{\beta_{\tau}^{\text{nonexp}}}{\beta_s^{\text{nonexp}}} g(\tau, X_{\tau}) \right] \\ &= \sup_{\tau \in [t, T]} \mathbb{E}_{t, x} \left[-\frac{\beta_{\tau}^{\text{nonexp}} (\beta_s^{\text{nonexp}})' }{(\beta_s^{\text{nonexp}})^2} g(\tau, X_{\tau}) \right] \\ &= -\frac{(\beta_s^{\text{nonexp}})'}{\beta_s^{\text{nonexp}}} \sup_{\tau \in [t, T]} \mathbb{E}_{t, x} \left[\frac{\beta_{\tau}^{\text{nonexp}}}{\beta_s^{\text{nonexp}}} g(\tau, X_{\tau}) \right] \\ &= -\frac{(\beta_s^{\text{nonexp}})'}{\beta_s^{\text{nonexp}}} f^K(t, x, s). \end{aligned}$$

Let $\gamma(t) := -\frac{(\beta_t^{\text{nonexp}})'}{\beta_t^{\text{nonexp}}}$ for $t \in [0, T]$. Then, $\frac{\partial f^K}{\partial s}(t, x, t) = -\frac{(\beta_t^{\text{nonexp}})'}{\beta_t^{\text{nonexp}}} V^K(t, x) = \gamma(t) V^K(t, x)$. Also note that, when $\beta_t = e^{-\rho t}$, $\gamma(t) = \rho$, which simplifies (A.49) to (A.48). So, the HJB equations (A.48) and (A.49) become

$$\text{exponential discounting: } \mathcal{L}V_{exp}^K - \rho V_{exp}^K + K(g - V_{exp}^K)^+ = 0 \quad (\text{A.52})$$

$$\text{non-exponential discounting: } \mathcal{L}V_{ne}^K - \gamma(t) V_{ne}^K + K(g - V_{ne}^K)^+ = 0. \quad (\text{A.53})$$

Define $W(t, x) = V_{ne}^K - V_{exp}^K$. Subtracting (A.52) from (A.53) yields a PDE for W ,

$$\mathcal{L}(V_{ne}^K - V_{exp}^K) - \gamma(t) V_{ne}^K + \rho V_{exp}^K + K[(g - V_{ne}^K)^+ - (g - V_{exp}^K)^+] \quad (\text{A.54})$$

$$= \mathcal{L}(W) - \gamma(t) W + (\rho - \gamma(t)) V_{exp}^K + K[(g - V_{ne}^K)^+ - (g - V_{exp}^K)^+] = 0, \quad (\text{A.55})$$

with terminal condition $W(T, x) = 0$.

From (9) in (Ebert et al., 2020) and under our requirement (5.41),

$$\rho = \int_0^\infty r dF(r) = -\frac{(\beta_0^{\text{nonexp}})'}{\beta_0^{\text{nonexp}}} = \gamma(0). \quad (\text{A.56})$$

We define the measures of decreasing impatience under β_t^{exp} and β_t^{nonexp} as

$$P_t^{\text{exp}} := -\frac{(\log(\beta_t^{\text{exp}}))''}{(\log(\beta_t^{\text{exp}}))'} \quad \text{and} \quad P_t^{\text{nonexp}} := -\frac{(\log(\beta_t^{\text{nonexp}}))''}{(\log(\beta_t^{\text{nonexp}}))'},$$

respectively. By Proposition 1 of (Ebert et al., 2020), all weighted discounting functions have non-negative measure of decreasing impatience, with the minimum attained by a simple exponential discounting function. So, we have, for any $t \in [0, T]$,

$$0 = P_t^{\text{exp}} \leq P_t^{\text{nonexp}} = -\frac{\left(\frac{(\beta_t^{\text{nonexp}})'}{\beta_t^{\text{nonexp}}}\right)'}{\left(\frac{(\beta_t^{\text{nonexp}})'}{\beta_t^{\text{nonexp}}}\right)},$$

which implies $\left(\frac{(\beta_t^{\text{nonexp}})'}{\beta_t^{\text{nonexp}}}\right)' \geq 0$. So, $\gamma(t)$ is decreasing in t and $\gamma(t) \leq \gamma(0) = \rho$ for all $t \in [0, T]$.

Suppose $W(t, x) < 0$ for some $(t, x) \in [0, T] \times \mathbb{R}_+$. Since $W(T, x) = 0$ and assuming W is smooth, the minimum of W is negative and is attained at an interior point, denoted by $(\hat{t}, \hat{x}) \in [0, T] \times \mathbb{R}_+$.

At $(\hat{t}, \hat{x})$, $W_x = 0$ and $W_{xx} \geq 0$ as it is the minimum. Moreover, $W_t \geq 0$ if $\hat{t} = 0$ otherwise $W_t = 0$. Therefore,

$$\mathcal{L}W(\hat{t}, \hat{x}) = W_t + \mu \hat{x} W_x + \frac{1}{2} \sigma^2 \hat{x}^2 W_{xx} \geq 0. \quad (\text{A.57})$$

Also, $-\gamma(\hat{t})W > 0$ since $(\beta_t^{\text{nonexp}})' < 0$ implies $\gamma(\hat{t}) > 0$. Moreover, $(\rho - \gamma(\hat{t}))V_{exp}^K \geq 0$ as $\gamma(\hat{t}) \leq \rho$ and $V_{exp}^K \geq 0$. Finally, $V_{ne}^K < V_{exp}^K$ at $(\hat{t}, \hat{x})$ implies $(g - V_{ne}^K)^+ - (g - V_{exp}^K)^+ \geq 0$.

As a result,

$$\mathcal{L}(W) - \gamma(\hat{t})W + (\rho - \gamma(\hat{t}))V_{exp}^K + K[(g - V_{ne}^K)^+ - (g - V_{exp}^K)^+] > 0, \quad (\text{A.58})$$

which contradicts the PDE of W (A.55), that the left hand side actually equals zero. Therefore, no such negative minimum W exists and hence $W \geq 0$, which implies $V_{ne}^K \geq V_{exp}^K$.

We note that this inequality holds for all choices of $K > 0$. According to Lemma 2.1 in (Dai et al., 2024), we further have V_{exp}^K and V_{ne}^K (the solutions to (A.48) and (A.49)) converges to V_{exp} and V_{ne} (the solutions to (5.42) and (5.43)) respectively as $K \rightarrow \infty$. In other words, V_{exp}, V_{ne} are respectively the solutions to (A.48) and (A.49) when $K \rightarrow \infty$. So, $V_{ne} \geq V_{exp}$.

A.4.4 Proof of Proposition (5.3.2)

For a given penalty parameter $K > 0$, recall that

$$\begin{aligned}\pi_{exp}^* &= \frac{1}{1 + \exp(\frac{1}{\lambda}K(V_{exp} - g))} = \frac{\exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))}, \\ \pi_{ne}^* &= \frac{1}{1 + \exp(\frac{1}{\lambda}K(V_{ne} - g))} = \frac{\exp(-\frac{1}{\lambda}K(V_{ne} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{ne} - g))}.\end{aligned}$$

Clearly, $V_{ne} \geq V_{exp}$ implies $\pi_{ne}^* \leq \pi_{exp}^*$. So, the remainder is to show that $V_{ne} \geq V_{exp}$.

For a given penalty parameter $K > 0$, denote by V_{exp}, V_{ne} the value functions under the exponential and non-exponential discounting setting, with exploratory stopping policy. Recall the HJB equation under exponential discounting and with exploratory policy (5.48), which we repeat here

$$\mathcal{L}V + \sup_{\pi \in [0,1]} \{-K\pi V + Kg(x)\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} - \rho V = 0.$$

Substituting $\pi = \pi_{exp}^* = \frac{\exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))}$ yields,

$$\begin{aligned}& \mathcal{L}V_{exp} - \frac{K(V_{exp} - g) \exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} - \rho V_{exp} \\& - \lambda \left\{ \frac{\exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \log \left(\frac{\exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right) \right. \\& \left. - \frac{\log(1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g)))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right\} \\& = \mathcal{L}V_{exp} - \frac{K(V_{exp} - g) \exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} - \rho V_{exp} \\& - \lambda \left\{ \frac{\exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \left[-\frac{1}{\lambda}K(V_{exp} - g) - \log(1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))) \right] \right. \\& \left. - \frac{\log(1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g)))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right\} \\& = \mathcal{L}V_{exp} - \frac{K(V_{exp} - g) \exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} - \rho V_{exp} + \frac{K(V_{exp} - g) \exp(-\frac{1}{\lambda}K(V_{exp} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))}\end{aligned}$$

$$+\lambda \log \left(1 + \exp\left(-\frac{1}{\lambda}K(V_{exp} - g)\right) \right) = 0,$$

which simplifies to

$$\mathcal{L}V_{exp} - \rho V_{exp} + \lambda \log \left(1 + \exp\left(-\frac{1}{\lambda}K(V_{exp} - g)\right) \right) = 0. \quad (\text{A.59})$$

Likewise, recall the HJB equation for the value function V_{ne} under non-exponential discounting with exploratory policy,

$$\mathcal{L}V + \sup_{\pi \in [0,1]} \{-K\pi V + Kg(x)\pi - \lambda(\pi \log \pi + (1 - \pi) \log(1 - \pi))\} - \gamma(t)V = 0, \quad (\text{A.60})$$

where we substitute $-\frac{\partial f}{\partial s}(t, x, t)$ with $-\gamma(t)V$, following the same steps as in the proof of Proposition 5.3.1; see details in Section A.4.3. Substituting $\pi = \pi_{ne}^* = \frac{\exp(-\frac{1}{\lambda}K(V_{ne}-g))}{1+\exp(-\frac{1}{\lambda}K(V_{ne}-g))}$ and after simplification,

$$\mathcal{L}V_{ne} - \gamma(t)V_{ne} + \lambda \log \left(1 + \exp\left(-\frac{1}{\lambda}K(V_{ne} - g)\right) \right) = 0. \quad (\text{A.61})$$

Subtracting (A.59) from (A.61) and defining $W = V_{ne} - V_{exp}$,

$$\begin{aligned} & \mathcal{L}(V_{ne} - V_{exp}) - \gamma(t)V_{ne} + \rho V_{exp} + \lambda \log \left(\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right) \\ &= \mathcal{L}(V_{ne} - V_{exp}) - \gamma(t)(V_{ne} - V_{exp}) + (\rho - \gamma(t))V_{exp} + \lambda \log \left(\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right) \\ &= \mathcal{L}W - \gamma(t)W + (\rho - \gamma(t))V_{exp} - KW + \lambda \log \left(\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right) = 0. \end{aligned} \quad (\text{A.62})$$

For contradiction, suppose $W < 0$, i.e., $V_{ne} < V_{exp}$, for some $(t, x) \in [0, T] \times \mathbb{R}$. Since $W(T, x) = 0$ and assuming W is smooth, the minimum of W is negative and is attained at an interior point, denoted by $(\hat{t}, \hat{x}) \in [0, T] \times \mathbb{R}_+$. Following the same argument as in Appendix A.4.3 (the proof of Proposition 5.3.1), we know that at $(\hat{t}, \hat{x})$, $\mathcal{L}W \geq 0$, $-\gamma(\hat{t})W > 0$, $(\rho - \gamma(\hat{t}))V_{exp} \geq 0$ and $-KW > 0$. Moreover, when $W < 0$, $\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne}-g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp}-g))} > 1$, implying $\log \left(\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne}-g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp}-g))} \right) > 0$. As a result,

$$\mathcal{L}W - \gamma(t)W + (\rho - \gamma(t))V_{exp} - KW + \lambda \log \left(\frac{1 + \exp(-\frac{1}{\lambda}K(V_{ne} - g))}{1 + \exp(-\frac{1}{\lambda}K(V_{exp} - g))} \right) > 0,$$

which contradicts the PDE for W (A.62). Therefore, no such negative minimum of W exists and hence $W \geq 0$, which implies $V_{ne} \geq V_{exp}$.

A.4.5 Proof of Theorem 5.4.1

Note that $\bar{J}_{ne}^\pi(t, x, r) = r \cdot J_{ne}^\pi(t, x) = r \cdot f_{ne}^\pi(t, x, t)$, for all $(t, x, r) \in [0, T] \times \mathbb{R} \times \mathbb{R}_+$. So, applying Itô's formula to (5.49),

$$\begin{aligned} dM_t^\pi &= d(R_t^\pi J_{ne}^\pi(t, X_t)) + R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] dt \\ &= d(R_t^\pi f^\pi(t, X_t, t)) + R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] dt \\ &= (f^\pi(t, X_t, t)) \cdot (-K\pi_t R_t^\pi) dt + R_t^\pi \cdot (df^\pi(t, X_t, t)) \\ &\quad + R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] dt. \end{aligned}$$

The drift term is

$$\begin{aligned} &(f^\pi(t, X_t, t)) \cdot (-K\pi_t R_t^\pi) + R_t^\pi \cdot \left(\mathcal{L}f^\pi(t, X_t, t) - \frac{\partial f^\pi}{\partial s}(t, X_t, t) \right) \\ &+ R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] \\ &= R_t^\pi \left[\mathcal{L}J_{ne}^\pi(t, X_t) - K\pi_t J_{ne}^\pi(t, X_t) + K\pi_t g(X_t) \right. \\ &\quad \left. - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t)) - \frac{\partial f^\pi}{\partial s}(t, X_t, t) \right], \end{aligned}$$

which equals zero if $\pi = \pi_{ne}^*$, by (5.26). So, M_t^π is a martingale when $\pi = \pi_{ne}^*$.

A.4.6 Proof of Corollary 5.4.1

Similar to (5.49), we write $\tilde{M}_t^\pi$

$$\tilde{M}_t^\pi = R_t^\pi \cdot J_{exp}^\pi(t, X_t) + \int_0^t R_v^\pi [Kg(X_v)\pi_v - \lambda(\pi_v \log \pi_v + (1 - \pi_v) \log(1 - \pi_v))] dv.$$

By Itô's formula,

$$\begin{aligned} d\tilde{M}_t^\pi &= d(R_t^\pi J_{exp}^\pi(t, X_t)) + R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] dt \\ &= (J_{exp}^\pi(t, X_t)) \cdot (-K\pi_t R_t^\pi) dt + R_t^\pi \cdot (dJ_{exp}^\pi(t, X_t)) \\ &\quad + R_t^\pi [Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] dt. \end{aligned}$$

The drift term is

$$(J_{exp}^\pi(t, X_t)) \cdot (-K\pi_t R_t^\pi) + R_t^\pi \cdot (\mathcal{L}J_{exp}^\pi(t, X_t) - \rho J_{exp}^\pi(t, X_t))$$

$$\begin{aligned}
& +R_t^\pi[Kg(X_t)\pi_t - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t))] \\
= & R_t^\pi \left[\mathcal{L}J_{exo}^\pi(t, X_t) - K\pi_t J_{ne}^\pi(t, X_t) + K\pi_t g(X_t) \right. \\
& \left. - \lambda(\pi_t \log \pi_t + (1 - \pi_t) \log(1 - \pi_t)) - \rho J_{exp}^\pi(t, X_t) \right],
\end{aligned}$$

which equals zero if $\pi = \pi_{exp}^*$, by (5.48).