Journal of Industrial and Management Optimization

https://www.aimspress.com/journal/jimo

JIMO, 22(2): 1087–1111.
DOI:10.3934/jimo.2026040
Received: 10 October 2025
Revised: 19 December 2025
Accepted: 12 January 2026
Published: 27 January 2026

Research article

Exploratory mean-variance portfolio selection with constant elasticity of variance models in regime-switching markets

Xiaoyu Xing*, Xingtian Zhang and Jiarou Luo

School of Sciences, Hebei University of Technology, Beichen District, Tianjin 300401, China

* **Correspondence:** Email: xingxiaoyutj@163.com.

Abstract: We studied a continuous-time exploratory mean-variance portfolio optimization problem using a reinforcement learning framework. The problem was set in a Markov regime-switching financial market, which captured time-varying market characteristics. Under different regimes, the risky asset followed constant elasticity of variance dynamics with regime-dependent parameters. Within this setting, the exploratory mean-variance portfolio optimization problem was formulated as a stochastic control problem. Stochastic dynamic programming techniques were employed to derive the Hamilton–Jacobi–Bellman equation associated with the exploratory control problem. We first derived analytical solutions for the optimal investment strategy and the corresponding value function. Although these solutions admitted closed-form representations, they were expressed in an integral form, which made them difficult to implement directly in practical numerical computations. Because of this, we developed a reinforcement learning algorithm to approximate the optimal investment policy and the corresponding value function. Based on the structural properties of the analytical solutions, we established the convergence of both the investment policy and the value function by invoking the policy improvement theorem. This result provided a rigorous theoretical foundation for the proposed algorithm. In the algorithmic implementation, linear function approximation was employed to parameterize both the value function and the investment policy. Finally, numerical experiments were conducted to verify the convergence behavior of the proposed algorithm and to demonstrate its effectiveness in solving the considered portfolio optimization problem.

Keywords: constant elasticity of variance; regime-switching markets; reinforcement learning; mean-variance portfolio selection; Choquet regularizer
Mathematics Subject Classification: 93E20, 93E35

1. Introduction

The mean-variance (MV) portfolio optimization in continuous time aims to maximize the investor's expected terminal wealth and minimize portfolio volatility at the fixed horizon. Markowitz [1] first proposed the MV portfolio optimization problem, but only for single-period cases. Since then, the problem has been extensively studied and extended. Li and Ng [2] investigated multiperiod portfolio problems. Numerous other extensions have followed. For instance, Zhou and Li [3] proposed stochastic control models of the linear–quadratic (LQ) type, and the MV efficient frontier was further analyzed by Chiu and Li [4]. Zhang et al. [5] derived equilibrium strategies via forward backward stochastic differential equations under the log-return based framework.

The classical mean-variance model mentioned above operates under the assumption that the financial market functions within a single regime. In reality, the market may exhibit distinct regimes with transitions among them (e.g., bear and bull phases in equity markets). Regime-switching dynamics are typically modeled by continuous-time Markov chains, where model parameters (e.g., equity returns, volatilities) vary according to the finite market regimes. Hamilton [6] pioneered the use of regime-switching models in the financial literature. Subsequent research extended Markov regime-switching models to option valuation, optimal asset allocation, and portfolio selection problems. Zhou and Yin [7] investigated a regime-switching framework in which the risk-free interest rate and stock volatility switch among finite states. Gal'perin et al. [8] studied a jump-diffusion risky asset model in which the drift, diffusion, and jump components are modulated by a homogeneous Markov chain. Chen et al. [9] integrated a Markov regime-switching model into asset selection framework. Xie [10], Chen and Yang [11] and Chen and Huang [12] examined an asset-liability model with Markov regime switching. Zhou [13] addressed an optimal investment-consumption model in which market dynamics are governed by a Markov regime-switching mechanism and the value function is approximated using a Markov chain approximation method. Eisenberg et al. [14] investigated reinsurance pricing by applying a two-state Markov regime-switching framework. Their work demonstrated that the solution to the Hamilton–Jacobi–Bellman (HJB) equation, as well as the resulting reinsurance strategy, can be uniquely represented through a recursive approach, with the strategy emerging as the limit of ordinary differential equation (ODE) solutions. Zhao and Song [15] explored liability valuation for life insurers in a regime-switching market, establishing an interest rate risk model with regime shifts and demonstrating its superior suitability for modeling interest rate risk.

The application of reinforcement learning (RL) is becoming increasingly prevalent in the financial field, with prominent use cases in algorithmic finance and quantitative investment strategies. While RL offers the advantage of learning autonomously without supervised data, it faces challenges like the exploration-exploitation trade-off, sparse rewards, and high-dimensional state spaces. Techniques such as experience replay, target networks, and entropy regularization help improve stability. In our paper, we use RL to solve a regime-switching MV problem, where an agent establishes an optimal policy by balancing exploration (testing new actions) and exploitation (capitalizing on known strategies) to maximize long-term risk-adjusted returns.

Wang et al. [16] introduced a theoretical framework that integrates entropy regularization into continuous-time relaxed control, providing a foundational analysis of the exploration–exploitation trade-off in RL. In their entropy-regularized framework, Wang and Zhou [17] developed an MV portfolio optimization model. Their analysis revealed that the optimal policy is Gaussian,

characterized by a variance that decays over time. Furthermore, they proposed an exploratory MV (EMV) algorithm and showed it achieved superior performance compared with traditional maximum likelihood estimation (MLE) as well as the deep deterministic policy gradient approach. Guo et al. [18] and Firoozi et al. [19] extended this RL framework to mean-field games. Jia and Zhou [20–22] introduced policy evaluation (PE)-based algorithms, policy gradient (PG)-based continuous-time algorithms, and Q-learning algorithms. Dai et al. [23] developed an extension of the exploratory stochastic control framework for incomplete markets characterized by regime-switching risky asset prices. For the time-inconsistent MV problem, the incorporation of an entropy regularizer induced exploration and led to a Nash equilibrium policy that followed a Gaussian distribution. Dai et al. [24] studied the Merton expected utility maximization problem in an incomplete market, proposing a recursive weighted exploration scheme that yields an optimal Gaussian policy.

Han et al. [25] introduced an alternative measure of behavioral randomness called Choquet regularization. They showed that, in infinite-horizon LQ control, the choice of regularizer dictates the form of the optimal exploration distribution, which is not necessarily Gaussian. Subsequently, they derived several explicit optimal distributions corresponding to different regularizers. Using Choquet regularization to quantify exploration, Guo et al. [26] extended this investigation to the exploratory MV problem. Guo [27] studied a competitive market involving two agents, modeled as a non-zero-sum differential game with partially and fully unknown model parameters. Under the RL framework, agents aim to maximize their own Choquet-regularized MV criterion. Their study continues to employ this regularization approach.

We extend the reinforcement learning framework to a financial market with regime switching. Most existing literature on regime switching focuses solely on parameter variations across different market states. Chen et al. [28] considered a risky asset modeled by a geometric Brownian motion (GBM) whose parameters depend on the market state. An MV portfolio selection problem featuring unobservable market regimes (bull/bear) was addressed by Wu and Li [29]. In this study, we first propose a regime-switching framework in which the risky asset is modeled via a constant elasticity of variance (CEV) process with state-dependent parameters, allowing for transitions between n distinct CEV models in different market states. This means the drift rate, volatility, and elasticity of variance coefficient can all vary with market regime shifts. Furthermore, this framework enables the investigation of the risky asset following distinct stochastic processes under different market regimes. For instance, it allows the risky asset to follow different stochastic volatility models in different states. This feature gives our research notable theoretical and practical significance. Second, inspired by Li et al. [30], we derive closed-form solutions for the optimal value function and policy under n distinct CEV models with regime-switching market states. Finally, we integrate the MV problem with RL techniques and develop a novel RL algorithm specifically designed for our proposed framework.

The following sections outline the content of this paper. Section 2 introduces an exploratory MV model with regime switching. We derive the optimal strategy solution in Section 3. Section 4 presents a policy iteration procedure along with its convergence proof. Section 5 proposes a RL algorithm. Section 6 exhibits the numerical results. Section 7 presents the concluding remarks, and the complete technical arguments together with auxiliary results are deferred to the Appendix.

2. Model formulation

Let $(\Omega, \mathbb{F}, P)$ be a complete probability space, where the filtration $\mathbb{F} := \{\mathcal{F}_t\}_{t\geq 0}$ is right-continuous, P-complete, and generated by a standard Brownian motion $\{W(t)\}_{t\geq 0}$ and a continuous-time homogeneous Markov process $\{\alpha_t\}_{t\geq 0}$. The investment horizon T is fixed, and the investor trades continuously over the period $[0, T]$.

We consider a financial market modulated by a continuous-time homogeneous Markov chain $\{\alpha_t\}_{t\geq 0}$ with a state space $\mathcal{N} = \{e_1, e_2, \ldots, e_n\}$, where $\{e_i\}_{1\leq i\leq n}$ represents the standard basis vector whose i-th component is 1 and all other components are 0. Here, $\alpha_t = e_i$ indicates that the market state at time t is in the i-th regime. The transition probabilities are defined as $p_{ij}(\tau) = p(\alpha_{t+\tau} = e_j|\alpha_t = e_i) = p(\alpha_\tau = e_j|\alpha_0 = e_i)$, $\forall t, \tau \geq 0, e_i, e_j \in \mathcal{N}$, and the transition probability matrix is denoted by $\mathbf{P}(t) = (p_{ij}(t))$. A transition intensity rate matrix $\mathbf{Q} = (q_{ij})_{n\times n} \in \mathbb{R}^{n\times n}$ can be derived from $\mathbf{P}(t)$ and usually used to generate the evolution of the Markov chain. Here, q_{ij} is the instantaneous transition intensity of the Markov chain from e_i to e_j. The entries q_{ij} satisfy: (I) $q_{ij} \geq 0$, if $i \neq j$; (II) $q_{ii} \leq 0$ and $q_{ii} = -\sum_{j\neq i} q_{ij}$; $i = 1, \cdots, n$. The process $\{\alpha_t\}_{t\geq 0}$ can be decomposed into the following semi-martingale representation

$$\alpha_t = \alpha_0 + \int_0^t \mathbf{Q}\alpha_u du + M(t), \tag{2.1}$$

where $M(t)$ is a martingale.

2.1. The classical regime-switching MV problem

We consider a financial market with two assets. The risk-free bond $S_0(t)$ satisfies

$$dS_0(t) = rS_0(t)dt, \tag{2.2}$$

where $r > 0$ represents the constant risk-free interest rate. The price process $S(t)$ of the risky asset follows a regime switching CEV model

$$\frac{dS(t)}{S(t)} = \mu(t, \alpha_t)dt + \sigma(t, \alpha_t)S(t)^{\beta(t,\alpha_t)}dW_t,\ S(0) = s_0 > 0, \tag{2.3}$$

where $\mu(t, \alpha_t)S(t)$ is the growth rate, and $\sigma(t, \alpha_t)S(t)^{\beta(t,\alpha_t)+1}$ is the volatility rate. $\beta(t, \alpha_t)$ is a constant elasticity coefficient, and it is generally assumed that $\beta(t, \alpha_t) \leq 0$. Here, the growth rate and the volatility rate depend on α_t.

Remark 2.1. *The variation in the elasticity coefficient across different regimes leads to model transitions among these regimes. For example, consider a two-regime system $\mathcal{N} = \{e_1, e_2\}$. When $\beta(t, e_1) = 0$, the risky asset model becomes $\frac{dS(t)}{S(t)} = \mu(t, e_1)dt + \sigma(t, e_1)dW(t)$, which follows a GBM. When $\beta(t, e_2) = -\frac{1}{2}$, the risky asset model becomes $\frac{dS(t)}{S(t)} = \mu(t, e_2)dt + \sigma(t, e_2)S(t)^{-\frac{1}{2}}dW(t)$, which follows a Cox-Ingersoll-Ross (CIR) process. Thus, the risky asset model dynamically switches between GBM and CIR process depending on the market regime.*

We consider a general situation in which the risky asset price model switches among n different CEV models. The investor allocates this portfolio between a risky asset and a risk-free asset. Let $u(t, \alpha_t)$ be the discounted value allocated to the risky asset at time t. The control process

$\{u = u(t,\alpha_t) : t \in [0,T], \alpha_t \in \mathcal{N}\}$ is adapted to $\{\mathcal{F}_t\}_{t\in[0,T]}$. Let $\{X^u(t,\alpha_t), 0 \le t \le T\}$ denote the investor's discounted wealth process. Under the self-financing strategy, $X^u(t,\alpha_t)$ satisfies

$$dX^u(t,\alpha_t) = (\mu(t,\alpha_t) - r)u(t,\alpha_t)dt + \sigma(t,\alpha_t)S(t)^{\beta(t,\alpha_t)}u(t,\alpha_t)dW_t, \tag{2.4}$$

with an initial wealth $X^u(t,\alpha_t) = x > 0$ and an initial regime $\alpha_t = e_i \in \mathcal{N}$. $u = \{u(t,\alpha_t) : t \in [0,T], \alpha_t \in \mathcal{N}\}$ is called an admissible strategy if

(i) $u(t,\alpha_t)$ is $\mathcal{F}_t$-progressively measurable;

(ii) $\forall t \in [0,T],\ E\left[\int_0^T (\sigma(t,\alpha_t)S(t)^{\beta(t,\alpha_t)}u(t,\alpha_t))^2 dt\right] < \infty$;

(iii) the pair $(u(t,\alpha_t), X^u(t,\alpha_t))$ is the unique strong solution to SDE (2.4) .

We aim to balance returns and mitigate risk by using the MV optimization framework on a fixed horizon $[0,T]$. Our goal is to minimize the variance of $X^u(T,\alpha_T)$ subject to a fixed expected value z. The problem is formulated as

$$\begin{aligned} &\min_u \mathrm{Var}[X^u(T,\alpha_T)], \\ &\text{subject to } E[X^u(T,\alpha_T)] = z. \end{aligned} \tag{2.5}$$

Since the objective function depends nonlinearly on terminal wealth, the MV problem exhibits time inconsistency. Generally, there are two approaches to solving this problem: solve the pre-commitment strategy [31] or solve the Nash equilibrium strategy [32]. We primarily seek the optimal pre-commitment strategy, where the investor determines the optimal strategy at the initial time and then consistently implements this predetermined strategy at all future moments. By introducing a Lagrange multiplier w, the constrained problem in (2.5) is transformed into a classical unconstrained problem

$$\min_u E\left(X^u(T,\alpha_T) - w\right)^2 - (w-z)^2. \tag{2.6}$$

The optimal solution to (2.6) is denoted by $u^* = \{u^*(t,\alpha_t), 0 \le t \le T\}$, which depends on w. The optimal Lagrange multiplier w^* is determined by solving $E\left[X^{u^*}(T,\alpha_T;w)\right] = z$.

Given a strategy u, we define the value function as

$$V^u(t,x,s,e_i) = E\left[\left(X^u(T,\alpha_T) - w\right)^2 \middle| X_t = x, S_t = s, \alpha_t = e_i\right] - (w-z)^2. \tag{2.7}$$

The classical optimal value function is defined as

$$V^{cl}(t,x,s,e_i) = \min_u E\left[\left(X^u(T,\alpha_T) - w\right)^2 \middle| X_t = x, S_t = s, \alpha_t = e_i\right] - (w-z)^2. \tag{2.8}$$

2.2. The EMV problem with regime switching

Reinforcement learning techniques do not require estimating model parameters. Instead, they learn optimal strategies through interaction with the market environment, achieved via a balance of exploration and exploitation. Building upon the framework of Wang et al. [16], we transform the control process in Equation (2.4) into a distributional control process. Consider $\Pi = \{\Pi_t(u), 0 \le t \le T\}$ as the probability distribution function of the strategy $u(t,\alpha_t)$, and let $\mathcal{M}$ represent the set of probability measures on $\mathbb{R}$. For any $\Pi \in \mathcal{M}$ and $x \in \mathbb{R}$, we have $\Pi(x) = \Pi((-\infty, x])$. Let $\mathcal{M}^p$ (where $p \in [1,\infty)$) denote the subset of $\mathcal{M}$ consisting of probability measures with finite p-th moments. If a random variable X has distribution Π, we write $X \sim \Pi$.

The probability distribution of $u(t,\alpha)$ is denoted by $\Pi \in \mathcal{M}^2$, with its mean and variance defined as $M(t,\alpha_t) := \int_{\mathbb{R}} u(t,\alpha_t)d\Pi_t(u)$ and $N^2(t,\alpha_t) := \int_{\mathbb{R}} u^2(t,\alpha_t)d\Pi_t(u) - M^2(t,\alpha_t)$. Then, equation (2.4) becomes

$$dX^{\Pi}(t) = (\mu(t,\alpha_t) - r)M(t,\alpha_t)dt + \sigma(t,\alpha_t)S_t^{\beta(t,\alpha_t)}\Big(M(t,\alpha_t)dW_t + N(t,\alpha_t)d\widetilde{W}_t\Big). \tag{2.9}$$

The derivation of the exploratory wealth process is provided in the Appendix. The Brownian motions $\widetilde{W}_t$ and W_t represent market noise sources. $\widetilde{W}_t$ specifically models exploration-induced noise which can be used for stochastic strategy generation. The coefficient of $d\widetilde{W}_t$ corresponds to the variance of Π_t, which quantifies the intensity of additional noise introduced into the system.

According to Han [25], to characterize the stochasticity of randomized strategies, we utilize the Choquet regularizer $\Phi_{\hat{h}}$. For a given bounded variation concave distortion function $\hat{h} : [0,1] \to \mathbb{R}$ with boundary conditions $\hat{h}(0) = 0$, $\hat{h}(1) = 1$ and any probability measure $\Pi \in \mathcal{M}$, the regularizer $\Phi_{\hat{h}}$ on $\mathcal{M}$ is defined as

$$\Phi_{\hat{h}}(\Pi) = \int_{\mathbb{R}} \hat{h} \circ \Pi([x,\infty))\,dx := \int_{-\infty}^{0} [\hat{h} \circ \Pi([x,\infty)) - \hat{h}(1)]\,dx + \int_{0}^{\infty} \hat{h} \circ \Pi([x,\infty))\,dx.$$

$\hat{h} \circ \Pi$ represents the distortion measure, and $\hat{h} \circ \Pi = \hat{h}(\Pi)$ denotes the composition of distortion function and probability measure. According to Wang et al. (Lemma 1) [33], the regularizer admits a representation based on quantile functions. For a distribution $\Pi \in \mathcal{M}$ and $p \in (0,1]$, the left quantile function (or lower quantile) is defined as $Q_{\Pi}(p) = \inf\{x \in \mathbb{R} : \Pi(x) \geqslant p\}$. When $\hat{h}$ is left-continuous, the regularizer admits the following quantile representation

$$\Phi_{\hat{h}}(\Pi) = \int_0^1 Q_{\Pi}(1-p)d\hat{h}(p).$$

The exploratory strategy means solving the EMV problem with regime switching within the RL framework. For any fixed w, we introduce an exploration weight $\lambda(t) > 0$ to derive the Choquet-regularized exploratory mean-variance problem with regime switching

$$\min_{\Pi \in \mathcal{A}(\Pi)} E\left[\Big(X^{\Pi}(T,\alpha_T) - w\Big)^2 - \int_t^T \lambda(\tau)\Phi_{\hat{h}}(\Pi)d\tau\right] - (w-z)^2. \tag{2.10}$$

The set $\mathcal{A}(\Pi)$ represents the admissible set of distributional controls, and the Lagrange multiplier can be determined via $E[X^{\Pi^*}(T,\alpha_T;w)] = z$.

Definition 2.1. *(admissible strategy) Let $\mathcal{B}(\mathbb{R})$ denote the Borel algebra on $\mathbb{R}$. The strategy $\Pi \in \mathcal{A}(\pi)$ is called an admissible strategy if*
(i) For $t \leqslant \tau \leqslant T$, it holds that $\Pi_\tau \in \mathcal{M}(\mathbb{R})$;
(ii) $\forall A \in \mathcal{B}(\mathbb{R})$, $\left\{\int_A \Pi_\tau(u)du, t \leqslant \tau \leqslant T\right\}$ is $\mathcal{F}_t$-progressively measurable;
(iii) $E\left[\int_t^T (\mu_\tau^2 + \sigma_\tau^2)d\tau\right] < \infty$;
(iv) $E\left[(X_T^{\Pi} - w)^2 - \int_t^T \lambda(\tau)\Phi_{\hat{h}}(\Pi_\tau)d\tau \middle| X_t^{\Pi} = x, S_t = s, \alpha_t = e_i\right] < \infty$.

Given a strategy Π , the value function is defined as

$$V^{\Pi}(t,x,s,e_i) = E\left[\Big(X^{\Pi}(T,\alpha_T) - w\Big)^2 - \int_t^T \lambda(\tau)\Phi_{\hat{h}}(\Pi)d\tau \middle| X_t^{\Pi} = x, S_t = s, \alpha_t = e_i\right] - (w-z)^2. \tag{2.11}$$

The optimal value function is defined as

$$V(t,x,s,e_i)=\min_{\Pi\in\mathcal{A}(\Pi)}E\left[\left(X^{\Pi}(T,\alpha_T)-w\right)^2-\int_t^T\lambda(\tau)\Phi_{\hat{h}}(\Pi)d\tau\middle|X_t^{\Pi}=x,S_t=s,\alpha_t=e_i\right]-(w-z)^2. \quad (2.12)$$

3. Solving the EMV problem with markov regime switching

To solve the EMV problem with regime switching, we apply the classical Bellman optimality principle

$$V(t,x,s,e_i)=\min_{\Pi\in\mathcal{A}(\Pi)}E\left[V(h,x,s,e_i)-\int_t^h\lambda(\tau)\Phi_{\hat{h}}(\Pi)d\tau\middle|X_t^{\Pi}=x,S_t=s,\alpha_t=e_i\right], \quad (3.1)$$

with $0\le t<h\le T,(t,x,s)\in[0,T]\times\mathbb{R}\times\mathbb{R}$. In brief, we denote $\mu_i=\mu(t,\alpha_t|\alpha_t=e_i)$, $\sigma_i=\sigma(t,\alpha_t|\alpha_t=e_i)$, $\beta_i=\beta(t,\alpha_t|\alpha_t=e_i)$, $M_i=M(t,\alpha_t|\alpha_t=e_i)$, $N_i=N(t,\alpha_t|\alpha_t=e_i)$. $V(t,s,x,e_i)$ satisfies the following HJB equation

$$\min_{\Pi\in\mathcal{M}(\Pi)}\left[(\mu_i-r)M_iV_x(t,s,x,e_i)+\frac{1}{2}\sigma_i^2s^{2\beta_i}\left(M_i^2+N_i^2\right)V_{xx}(t,s,x,e_i)+\sigma_i^2s^{2\beta_i+1}M_iV_{sx}(t,s,x,e_i)-\lambda(t)\Phi_{\hat{h}}(\Pi)\right]$$
$$+V_t(t,s,x,e_i)+\frac{1}{2}\sigma_i^2s^{2\beta_i+2}V_{ss}(t,s,x,e_i)+\mu_isV_s(t,s,x,e_i)+\sum_{j=1}^{n}q_{ij}V\left(t,x,s,e_j\right)=0,$$

$i=1,2,\ldots,n$. The terminal condition is $V(T,x,s,e_i)=(x-w)^2-(w-z)^2$. The minimization term in the above equation is denoted as

$$\varphi(t,x,s,e_i,\Pi)=(\mu_i-r)M_iV_x(t,s,x,e_i)+\frac{1}{2}\sigma_i^2s^{2\beta_i}\left(M_i^2+N_i^2\right)V_{xx}(t,s,x,e_i)+\mu_isV_s(t,s,x,e_i)$$
$$+\sigma_i^2s^{2\beta_i+1}M_iV_{sx}(t,s,x,e_i)-\lambda(t)\Phi_{\hat{h}}(\Pi).$$

We can observe that the function $\varphi(t,x,s,e_i,\Pi)$ depends solely on M_i and N_i^2 besides $\Phi_{\hat{h}}(\Pi)$. Then, we have

$$\min_{\Pi\in\mathcal{M}(\Pi)}\varphi(t,x,s,e_i,\Pi)=\min_{m_i\in\mathbb{R},n_i>0}\ \min_{\substack{\Pi\in\mathcal{M}(\mathbb{R})\\ \mu(\Pi)=m_i,\sigma(\Pi)^2=n_i^2}}\varphi(t,x,s,e_i,\Pi).$$

The embedded minimization problem equivalently becomes

$$\max_{\Pi\in\mathcal{M}(R)}\Phi_{\hat{h}}(\Pi)\ \text{ s.t. } M_i=m_i,N_i^2=n_i^2.$$

We need the following lemma given by Han et al. (2023) [25].

Lemma 1. *If $\hat{h}$ is continuous and not identically zero, then the optimization problem*

$$\max_{\Pi\in\mathcal{M}^2}\Phi_{\hat{h}}(\Pi)\quad s.t.\quad \mu(\Pi)=\mu_t \text{ and } \sigma^2(\Pi)=\sigma_t^2. \quad (3.2)$$

The quantile function of the optimal control Π^* satisfies

$$Q_{\Pi^*}(p)=\mu_t+\sigma_t\frac{\hat{h}'(1-p)}{\|\hat{h}'\|_2},\quad a.e.\quad for\ p\in(0,1), \quad (3.3)$$

and the maximum value of (3.2) *is* $\Phi_{\hat{h}}(\Pi^*) = \sigma_t\|\hat{h}'\|_2$*, where* $\hat{h}'$ *is the right-hand derivative of* $\hat{h}$ *and* $\|\hat{h}'\|_2 = \left(\int_0^1 (\hat{h}'(p))^2 dp\right)^{1/2}$.

According to Lemma 1, the quantile function $Q_{\Pi^*}(p)$ of the optimal strategy Π^* satisfies

$$Q_{\Pi^*}(p) = M_i + N_i \frac{\hat{h}'(1-p)}{\|\hat{h}'\|_2},$$

and $\Phi_{\hat{h}}(\Pi^*) = N_i\|\hat{h}'\|_2$, thus, the mean and variance can be expressed as

$$\begin{aligned}(M_i^*, N_i^*) = \arg\min_{m_i\in\mathbb{R}, n_i>0}\Big[&(\mu_i - r)m_i V_x(t,s,x,e_i) + \frac{1}{2}\sigma_i^2 s^{2\beta_i}\left(m_i{}^2 + n_i{}^2\right)V_{xx}(t,s,x,e_i)\\ &+\sigma_i^2 s^{2\beta_i+1} m_i V_{sx}(t,s,x,e_i) - \lambda(t)n_i\|\hat{h}'\|_2\Big] + V_t(t,s,x,e_i)\\ &+\frac{1}{2}\sigma_i^2 s^{2\beta_i+2}V_{ss}(t,s,x,e_i) + \mu_i s V_s(t,s,x,e_i) + \sum_{j=1}^{n} q_{ij}V\left(t,x,s,e_j\right).\end{aligned}$$

The first-order condition yields the following expressions for the optimal mean and standard deviation

$$M_i^* = -\frac{(\mu_i - r)V_x(t,s,x,e_i) + \sigma_i^2 s^{2\beta_i+1}V_{sx}(t,s,x,e_i)}{\sigma_i^2 s^{2\beta_i}V_{xx}(t,s,x,e_i)}, \tag{3.4}$$

$$N_i^* = \frac{\lambda(t)\left\|\hat{h}'\right\|_2}{\sigma_i^2 s^{2\beta_i}V_{xx}(t,s,x,e_i)}. \tag{3.5}$$

We conjecture that the value function is

$$V(t,s,x,e_i) = a(t,s,e_i)(x-w)^2 - (w-z)^2 + b(t,s,e_i). \tag{3.6}$$

Substituting the optimal strategy (3.4) (3.5) and the value function (3.6) into the HJB equation yields

$$\begin{aligned}&a_t(t,s,e_i)(x-w)^2 + b_t(t,s,e_i) + (\mu_i - r)\left(-\frac{(\mu_i - r)2a(t,s,e_i) + \sigma_i^2 s^{2\beta_i+1}2a_s(t,s,e_i)}{\sigma_i^2 s^{2\beta_i}2a(t,s,e_i)}\right)2a(t,s,e_i)(x-w)^2\\ &+\frac{1}{2}\sigma_i^2 s^{2\beta_i}\left[\left(\frac{(\mu_i - r)2a(x-w) + \sigma_i^2 s^{2\beta_i+1}2a_s(t,s,e_i)(x-w)}{\sigma_i^2 s^{2\beta_i}2a(t,s,e_i)}\right)^2 + \left(\frac{\lambda(t)\left\|\hat{h}'\right\|_2}{2\sigma_i^2 s^{2\beta_i}a(t,s,e_i)}\right)^2\right]2a(t,s,e_i)\\ &+\mu_i s\left(a_s(t,s,e_i)(x-w)^2 + b_s(t,s,e_i)\right) + \frac{1}{2}\sigma_i^2 s^{2\beta_i+2}\left(a_{ss}(t,s,e_i)(x-w)^2 + b_{ss}(t,s,e_i)\right)\\ &+\sigma_i^2 s^{2\beta_i+1}\left(-\frac{(\mu_i - r)2(x-w)a(t,s,e_i) + 2\sigma_i^2 s^{2\beta_i+1}a_s(t,s,e_i)(x-w)}{2\sigma_i^2 s^{2\beta_i}a(t,s,e_i)}\right)(2a_s(t,s,e_i)(x-w))\\ &-\frac{\lambda^2(t)\left\|\hat{h}'\right\|_2^2}{2\sigma_i^2 s^{2\beta_i}a(t,s,e_i)} + \sum_{j=1}^{n} q_{ij}\left(a\left(t,s,e_j\right)(x-w)^2 - (w-z)^2 + b\left(t,s,e_j\right)\right) = 0.\end{aligned}$$

By variable separation, we can obtain $a(t,s,e_i)$, $b(t,s,e_i)$ should satisfy the following equations

$$\begin{aligned} &a_t(t,s,e_i)-\left(\frac{(\mu_i-r)^2}{\sigma_i^2 s^{2\beta_i}}\right)a(t,s,e_i)+(-2(\mu_i-r)+\mu_i)sa_s(t,s,e_i)-\sigma_i^2 s^{2\beta+2}\frac{a_s(t,s,e_i)^2}{a(t,s,e_i)}\\ &+\frac{1}{2}\sigma_i^2 s^{2\beta_i+2}a_{ss}(t,s,e_i)+\sum_{j=1}^{n}q_{ij}a\left(t,s,e_j\right)=0, \end{aligned} \tag{3.7}$$

$$b_t(t,s,e_i)-\frac{1}{4}\frac{\lambda^2(t)\left\|\hat{h}'\right\|_2^2}{\sigma_i^2 s^{2\beta_i}a(t,s,e_i)}+\mu_i sb_s(t,s,e_i)+\frac{1}{2}\sigma_i^2 s^{2\beta_i+2}b_{ss}(t,s,e_i)+\sum_{j=1}^{n}q_{ij}b\left(t,s,e_j\right)=0, \tag{3.8}$$

and satisfy the terminal condition $a(T,s,e_i)=1,\ b(T,s,e_i)=0,\ i=1,\cdots,n$. Subsequently, we will evaluate for the parameters $a(t,s,e_i)$ and $b(t,s,e_i)$.

Proposition 3.1. *Let* $\mathbf{a}(t,s)=[a(t,s,e_1),\cdots,a(t,s,e_n)]^\top$, *then*

$$\mathbf{a}(t,s)=\mathbf{H}(t)\exp(\mathbf{K}(t)\mathbf{Y}), \tag{3.9}$$

where vector $\exp(\mathbf{K}(t)\mathbf{Y})$ *represents applying the exponential function element-wise to each element of vector* $\mathbf{K}(t)\mathbf{Y}$, $\mathbf{H}(t)$ *and* $\mathbf{K}(t)$ *satisfy* (3.16) *and* (3.17).

Proof. Since (3.7) is a nonlinear second-order partial differential equation whose explicit solution is difficult to obtain directly, we can nevertheless employ a power transformation and change of variables to convert this nonlinear PDE into a linear partial differential equation. Let

$$a(t,s,e_i)=f(t,y_i,e_i)\qquad y_i=s^{-2\beta_i},$$

with the terminal conditions $f(T,y,e_i)=1$, then we have

$$a_t(t,s,e_i)=f_t(t,y_i,e_i),\ a_s(t,s,e_i)=-2\beta_i f_{y_i}(t,y_i,e_i)s^{-2\beta_i-1},$$

$$a_{ss}(t,s,e_i)=-2\beta_i\left[-s^{-2\beta_i-2}(2\beta_i+1)f_{y_i}(t,y_i,e_i)-2\beta_i s^{-4\beta_i-2}f_{y_iy_i}(t,y_i,e_i)\right].$$

Substituting these partial derivatives into Equation (3.7), we obtain

$$\begin{aligned} &f_t(t,y_i,e_i)-\frac{(\mu_i-r)^2}{\sigma_i^2}y_i f(t,y_i,e_i)-(2r-\mu_i)2\beta_i y_i f_{y_i}(t,y_i,e_i)+\sigma_i^2\beta_i(2\beta_i+1)f_{y_i}(t,y_i,e_i)\\ &+2\sigma_i^2\beta_i^2 y_i f_{y_iy_i}(t,y_i,e_i)-4\sigma_i^2\beta_i^2\frac{f_{y_i}^2(t,y_i,e_i)}{f(t,y_i,e_i)}y_i+\sum_{j=1}^{n}q_{ij}f\left(t,y_j,e_j\right)=0. \end{aligned} \tag{3.10}$$

For convenience, we define some new notations

$$\begin{aligned} &\mathbf{A}=\mathrm{diag}\left(-\frac{(\mu_1-r)^2}{\sigma_1^2},\cdots,-\frac{(\mu_n-r)^2}{\sigma_n^2}\right), &&\mathbf{B}=\mathrm{diag}(-2(2r-\mu_1)\beta_1,\cdots,-2(2r-\mu_n)\beta_n),\\ &\mathbf{C}=\mathrm{diag}(\sigma_1^2\beta_1(2\beta_1+1),\cdots,\sigma_n^2\beta_n(2\beta_n+1)), &&\mathbf{D}=\mathrm{diag}(-4\sigma_1^2\beta_1^2,\cdots,-4\sigma_n^2\beta_n^2),\\ &\mathbf{E}=\mathrm{diag}(2\sigma_1^2\beta_1^2,\cdots,2\sigma_n^2\beta_n^2), &&\mathbf{f}(t,\mathbf{Y})=[f(t,y_1,e_1),\cdots,f(t,y_n,e_n)]^\top,\\ &\mathbf{Y}=[y_1,\cdots,y_n]^\top=[s^{-2\beta_1},\cdots,s^{-2\beta_n}]^\top, &&\widetilde{\mathbf{Y}}=\mathrm{diag}(\mathbf{Y}). \end{aligned}$$

We define the notations, then Equation (3.8) can be transformed into the following matrix equation

$$\frac{\partial \mathbf{f}}{\partial t} + \mathbf{A}\widetilde{\mathbf{Y}}\mathbf{f} + \mathbf{B}\widetilde{\mathbf{Y}}\nabla_{\mathbf{Y}}\mathbf{f} + \mathbf{C}\nabla_{\mathbf{Y}}\mathbf{f} + \mathbf{E}\widetilde{\mathbf{Y}}\Delta_{\mathbf{Y}}\mathbf{f} + \mathbf{D}\widetilde{\mathbf{Y}}\left(\frac{(\nabla_{\mathbf{Y}}\mathbf{f})^2}{\mathbf{f}}\right) + \mathbf{Q}\mathbf{f} = \mathbf{0}, \tag{3.11}$$

where $\nabla_{\mathbf{Y}}\mathbf{f}$ and $\Delta_{\mathbf{Y}}\mathbf{f}$ represent the first and second-order derivatives of the vector $\mathbf{f}$ with respect to the vector $\mathbf{Y}$ in element-wise sense. $(\nabla_{\mathbf{Y}}\mathbf{f})^2$ represents the element-wise multiplication of the vector $\nabla_{\mathbf{Y}}\mathbf{f}$ with itself. $\frac{(\nabla_{\mathbf{Y}}\mathbf{f})^2}{\mathbf{f}}$ represents the element-wise division of the vector $(\nabla_{\mathbf{Y}}\mathbf{f})^2$ by the vector $\mathbf{f}$. For the differential equation (3.11), we assume a solution of the form

$$\mathbf{f}(t, \mathbf{Y}) = \mathbf{H}(t)\exp(\mathbf{K}(t)\mathbf{Y}), \tag{3.12}$$

where

$$\mathbf{H}(t) = \operatorname{diag}(h_1(t), ..., h_n(t)), \mathbf{K}(t) = \operatorname{diag}(k_1(t), ..., k_n(t)).$$

$\mathbf{H}(t)$ and $\mathbf{K}(t)$ satisfy the terminal conditions $\mathbf{H}(T) = \mathbf{I}$, and $\mathbf{K}(T) = \mathbf{0}$, where $\mathbf{0}$ and $\mathbf{I}$ represent the $n \times n$ zero matrix and identity matrix. Notice that $\nabla_{\mathbf{Y}}\mathbf{f} = \mathbf{K}(t)\mathbf{f}$, $\Delta_{\mathbf{Y}}\mathbf{f} = \mathbf{K}(t)^2\mathbf{f}$, $\frac{(\nabla_{\mathbf{Y}}\mathbf{f})^2}{\mathbf{f}} = \mathbf{K}(t)^2\mathbf{f}$. Substituting (3.12) into (3.11), we get

$$\begin{aligned} &\dot{\mathbf{H}}(t)e^{\mathbf{KY}} + \mathbf{H}(t)\dot{\mathbf{K}}(t)\widetilde{\mathbf{Y}}e^{\mathbf{KY}} + \mathbf{A}\widetilde{\mathbf{Y}}\mathbf{H}(t)e^{\mathbf{KY}} + \mathbf{B}\widetilde{\mathbf{Y}}\mathbf{K}(t)\mathbf{H}(t)e^{\mathbf{KY}} + \mathbf{C}\mathbf{K}(t)\mathbf{H}(t)e^{\mathbf{KY}} \\ &+ \mathbf{E}\widetilde{\mathbf{Y}}\mathbf{K}(t)^2\mathbf{H}(t)e^{\mathbf{KY}} + \mathbf{D}\widetilde{\mathbf{Y}}\mathbf{K}(t)^2\mathbf{H}(t)e^{\mathbf{KY}} + \mathbf{Q}\mathbf{H}(t)e^{\mathbf{KY}} = \mathbf{0}. \end{aligned} \tag{3.13}$$

Applying separation of variables yields

$$\dot{\mathbf{H}}(t) + \mathbf{C}\mathbf{K}(t)\mathbf{H}(t) + \mathbf{Q}\mathbf{H}(t) = \mathbf{0}, \tag{3.14}$$

$$\dot{\mathbf{K}}(t) + \mathbf{A} + \mathbf{B}\mathbf{K}(t) - \mathbf{E}\mathbf{K}^2(t) = \mathbf{0}. \tag{3.15}$$

Subject to the terminal conditions $\mathbf{H}(T) = \mathbf{I}$ and $\mathbf{K}(T) = \mathbf{0}$, the solutions of the aforementioned differential equations are given by

$$\mathbf{H}(t) = \exp\left(\int_t^T (\mathbf{Q} + \mathbf{C}\mathbf{K}(\tau))d\tau\right), \quad \mathbf{K}(t) = \operatorname{diag}(k_1(t), \cdots, k_n(t)), \tag{3.16}$$

where k_i satisfies

(a) when $\beta_i \neq 0$,

$$\text{if } \mu_i = \sqrt{2}r, \text{ then } k_i(t) = \frac{r^2(2-\sqrt{2})^2(T-t)}{2\sigma_i^2\beta_i\left(1+r(2-\sqrt{2})\beta_i(T-t)\right)},$$

$$\text{if } \mu_i > \sqrt{2}r, \text{ then } k_i(t) = \frac{-(2r-\mu_i)+\sqrt{\mu_i^2-2r^2}\tan\left(\beta_i\sqrt{\mu_i^2-2r^2}(t-T))+\arctan\left(\frac{2r-\mu_i}{\sqrt{\mu_i^2-2r^2}}\right)\right)}{2\sigma_i^2\beta_i},$$

$$\text{if } \mu_i < \sqrt{2}r, \text{ then } k_i(t) = \frac{-(2r-\mu_i+\sqrt{2r^2-\mu_i^2})\sqrt{2r^2-\mu_i^2}}{\sigma_i^2\beta_i\left((2r-\mu_i-\sqrt{2r^2-\mu_i^2})e^{2\beta_i\sqrt{2r^2-\mu_i^2}(t-T)})-(2r-\mu_i+\sqrt{2r^2-\mu_i^2})\right)} - \frac{2r-\mu_i+\sqrt{2r^2-\mu_i^2}}{2\sigma_i^2\beta_i},$$

(b) when $\beta_1 = 0$, then $k_i(t) = \frac{(\mu_i-r)^2}{\sigma_i^2}(t-T)$. (3.17)

Proposition 3.2. *For $\alpha_\tau = e_m$, we denote $\beta_m = \beta(\tau, e_m)$, $\sigma_m = \sigma(\tau, e_m)$, $k_m = k(\tau, e_m)$, $h_m = h(\tau, e_m)$. The expression for $b(t, s, e_i)$ is*

$$b(t,s,e_i) = -\int_t^T \sum_{m=1}^n p_{im}(\tau-t)\frac{1}{4}\frac{\lambda^2(\tau)\|\hat{h}'\|_2^2}{\sigma_m^2}E\left(S_\tau^{-2\beta_m}a^{-1}(\tau,S_\tau,e_m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right)d\tau, \quad (3.18)$$

with

$$E\left(S_\tau^{-2\beta_m}a^{-1}(\tau,S_\tau,m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right) = \begin{cases} h_m^{-1}e^{-k_m}, & \text{if } \beta_m = 0, \\ h_m^{-1}4\beta_m^2\left(-\frac{\partial}{\partial\eta}F\left(\eta,\tau-t,y_0|\eta=\hat{k}_m,\ y_0=\frac{1}{4\beta_i^2}s^{-2\beta_i}\right)\right), & \text{if } \beta_m < 0, \end{cases}$$

where $\frac{\partial}{\partial\eta}F(\eta,\tau-t,y_0)$ satisfies (3.25)*, and the transition probability satisfies $p_{im}(\tau-t) = p(\alpha_\tau = e_m|\alpha_t = e_i)$.*

Proof. From Equation (2.3), S_τ satisfies $dS_\tau = \mu(\tau,\alpha_\tau)S_\tau d\tau + \sigma(\tau,\alpha_\tau)S_\tau^{\beta(\tau,\alpha_\tau)+1}dW_\tau$, $t \le \tau \le T$. By Itô's formula for $b(\tau, S_\tau, \alpha_\tau)$

$$db(\tau,S_\tau,\alpha_\tau) = \left(b_\tau + \mu(\tau,\alpha_\tau)S_\tau b_s + \frac{1}{2}\sigma(\tau,\alpha_\tau)^2S_\tau^{2\beta(\tau,\alpha_\tau)+2}b_{ss} + \sum_{j=1}^n q_{\alpha_\tau j}b\left(\tau,S_\tau,e_j\right)\right)d\tau + \sigma(\tau,\alpha_\tau)S_\tau^{\beta(\tau,\alpha_\tau)+1}b_s dW_\tau.$$

Taking the integral of both sides over $[t,T]$, we obtain

$$\int_t^T db(\tau,S_\tau,\alpha_\tau)=\int_t^T\left[b_\tau+\mu(\tau,\alpha_\tau)S_\tau b_s+\frac{1}{2}\sigma(\tau,\alpha_\tau)^2S_\tau^{2\beta(\tau,\alpha_\tau)+2}b_{ss}+\sum_{j=1}^n q_{\alpha_\tau j}b\left(\tau,S_\tau,e_j\right)\right]d\tau$$
$$+\int_t^T\sigma(\tau,\alpha_\tau)S_\tau^{\beta(\tau,\alpha_\tau)+1}b_s dW_\tau.$$

Considering that $b(T,S_T,\alpha_T)=0$ and $S_t=s,\ \alpha_t=e_i$, then

$$\int_t^T db(\tau,S_\tau,\alpha_\tau)=-b(t,s,e_i).$$

Taking the conditional expectation operator given $(S_t,\alpha_t)=(s,e_i)$, the martingale term vanishes, then

$$-b(t,s,e_i)=E\left[\int_t^T\left[b_\tau+\mu(\tau,\alpha_\tau)S_\tau b_s+\frac{1}{2}\sigma(\tau,\alpha_\tau)^2S_\tau^{2\beta(\tau,\alpha_\tau)+2}b_{ss}+\sum_{j=1}^n q_{\alpha_\tau j}b\left(\tau,S_\tau,e_j\right)\right]d\tau\middle|S_t=s,\ \alpha_t=e_i\right].$$

Let $g(\tau,S_\tau,\alpha_\tau)=-\frac{1}{4}\frac{\lambda^2(\tau)\|\hat{h}'\|_2^2}{\sigma^2(\tau,\alpha_\tau)S_\tau^{2\beta(\tau,\alpha_\tau)}a(\tau,S_\tau,\alpha_\tau)}$. Recall Equation (3.8) that can be simplified to

$$\begin{aligned}b(t,s,e_i)&=E\left(\int_t^T g(\tau,S_\tau,\alpha_\tau)\,d\tau\middle|S_t=s,\ \alpha_t=e_i\right)\\&=-\int_t^T\sum_{m=1}^n p_{im}(\tau-t)\frac{1}{4}\frac{\lambda^2(\tau)\|\hat{h}'\|_2^2}{\sigma^2(\tau,e_m)}E\left(S_\tau^{-2\beta(\tau,e_m)}a^{-1}(\tau,S_\tau,e_m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right)d\tau.\end{aligned}\tag{3.19}$$

In the following, we compute $E\left(S_\tau^{-2\beta(\tau,e_m)}a^{-1}(\tau,S_\tau,e_m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right)$, indeed

$$\begin{aligned}&E\left(S_\tau^{-2\beta(\tau,e_m)}a^{-1}(\tau,S_\tau,e_m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right)\\=&E\left(S_\tau^{-2\beta(\tau,e_m)}f^{-1}(\tau,S_\tau^{-2\beta(\tau,e_m)},e_m)\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right)\\=&h^{-1}(\tau,e_m)E\left(S_\tau^{-2\beta(\tau,e_m)}e^{-k(\tau,e_m)S_\tau^{-2\beta(\tau,e_m)}}\middle|S_t=s,\alpha_t=e_i,\alpha_\tau=e_m\right).\end{aligned}\tag{3.20}$$

(i) When $\beta(\tau,e_m)=0$, the integral simplifies to

$$b(t,s,e_i)=-\int_t^T\sum_{m=1}^n p_{im}(\tau-t)\frac{1}{4}\frac{\lambda^2(\tau)\|\hat{h}'\|_2^2}{\sigma^2(\tau,e_m)}h^{-1}(\tau,e_m))e^{-k(\tau,e_m)}d\tau,\tag{3.21}$$

where $h(\tau,e_m)$ and $k(\tau,e_m)$ are determined by equations (3.16) and (3.17).
(ii) When $\beta(\tau,e_m)<0$. We consider a CIR process which is constructed by $Y(\tau,e_m)=\frac{1}{4\beta(\tau,e_m)^2}S_\tau^{-2\beta(\tau,e_m)}$, by Itô's formula

$$dY(\tau,e_m)=\gamma(\tau,e_m)\left(\theta(\tau,e_m)-Y(\tau,e_m)\right)d\tau+\hat{\sigma}(\tau,e_m)\sqrt{Y(\tau,e_m)}dW_\tau,\tag{3.22}$$

with parameters $\gamma(\tau,e_m)=2\mu(\tau,e_m)\beta(\tau,e_m)$, $\theta(\tau,e_m)=\sigma^2(\tau,e_m)\frac{2\beta(\tau,e_m)+1}{4\gamma(\tau,e_m)\beta(\tau,e_m)}$, $\hat{\sigma}(\tau,e_m)=-\mathrm{sgn}(\beta(\tau,e_m))\sigma(\tau,e_m)$, $\gamma(\tau,e_m)\theta(\tau,e_m)=\sigma^2(\tau,e_m)\frac{2\beta(\tau,e_m)+1}{4\beta(\tau,e_m)}$. Denote $\hat{\sigma}_m=\hat{\sigma}(\tau,e_m)$,

$\gamma_m = \gamma(\tau, e_m)$, $\beta_m = \beta(\tau, e_m)$, $\sigma_m = \sigma(\tau, e_m)$, $\theta_m = \theta(\tau, e_m)$, $\hat{k}_m = \hat{k}(\tau, e_m)$, $k_m = k(\tau, e_m)$, $Y_m = Y(\tau, e_m)$. We can get a simpler expression for (3.20)

$$\begin{aligned} &h^{-1}(\tau, e_m) E\left(S_\tau^{-2\beta_m} e^{-k_m S_\tau^{-2\beta_m}} \middle| S_t = s, \alpha_t = e_i, \alpha_\tau = e_m\right) \\ &= h^{-1}(\tau, e_m) 4\beta_m^2 E\left(Y_m e^{-\hat{k}_m Y_m} \middle| S_t = s, \alpha_t = e_i, \alpha_\tau = e_m\right), \end{aligned} \tag{3.23}$$

where $\hat{k}_m = 4k_m\beta_m^2$. Now $E(Y_m e^{-\hat{k}_m Y_m} \big| S_t = s, \alpha_t = e_i, \alpha_\tau = e_m)$ will be determined, for this, let $F(\eta, \tau - t, y_0) = E(e^{-\eta Y_m} \big| Y_t = y_0)$, where the initial value of process Y_m is $y_0 = \frac{1}{4\beta_i^2} s^{-2\beta_i}$, then by the Laplace transform

$$E(Y_m e^{-\hat{k}_m Y_m} \big| S_t = s, \alpha_t = e_i, \alpha_\tau = e_m) = -\frac{\partial}{\partial \eta} F\left(\eta, \tau - t, y_0 | \eta = \hat{k}_m, y_0 = \frac{1}{4\beta_i^2} s^{-2\beta_i}\right).$$

According to Jeanblanc et al.(2009) [34]

$$F(\eta, \tau - t, y_0) = \exp\left\{-A_\eta(\tau - t) - y_0 G_\eta(\tau - t)\right\}, \tag{3.24}$$

with

$$\begin{aligned} G_\eta(\tau - t) &= \frac{2\gamma_m \eta}{\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}}, \\ A_\eta(\tau - t) &= -\frac{2\gamma_m \theta_m}{\hat{\sigma}_m^2} \ln \frac{2\gamma_m e^{\gamma_m(\tau - t)}}{\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}}. \end{aligned}$$

The partial derivative is derived as

$$\begin{aligned} \frac{\partial F}{\partial \eta}(\eta, \tau - t, y_0) &= -\left(A'_\eta(\tau - t) + y_0 G'_\eta(\tau - t)\right) \exp\left(-A_\eta(\tau - t) - y_0 G_\eta(\tau - t)\right) \\ &= -\left[\frac{2\gamma_m \theta_m (e^{\gamma_m(\tau - t)} - 1)}{\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}} + y_0 \frac{4\gamma_m^2 e^{\gamma_m(\tau - t)}}{\left[\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}\right]^2}\right] \\ &\quad \times \exp\left[\frac{2\gamma_m \theta_m}{\hat{\sigma}_m^2} \ln \frac{2\gamma_m e^{\gamma_m(\tau - t)}}{\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}} - y_0 \frac{2\gamma_m \eta}{\hat{\sigma}_m^2 \eta (e^{\gamma_m(\tau - t)} - 1) + 2\gamma_m e^{\gamma_m(\tau - t)}}\right]. \end{aligned} \tag{3.25}$$

Theorem 3.1. *Denote* $\mathbf{b}(t, s) = [b(t, s, e_1), \cdots, b(t, s, e_n)]^\top$. $\mathbf{a}(t, s)$ *and* $\mathbf{b}(t, s)$ *are determined by equations* (3.9), (3.18). $\mathbf{1}_n^\top$ *represents a column vector where all elements are equal to 1. Define the value function vector under different regimes as* $\mathbf{V}(t, s, x) = [V(t, s, x, e_1), \cdots, V(t, s, x, e_n)]^\top$, *then* $\mathbf{V}(t, s, x)$ *can be expressed as follows*

$$\mathbf{V}(t, s, x) = \mathbf{a}(t, s)(x - w)^2 - (w - z)^2 \mathbf{1}_n^\top + \mathbf{b}(t, s), \tag{3.26}$$

The optimal strategy associated with regime i is denoted by Π_i^*, *with its quantile function given by*

$$\mathbf{Q}_{\Pi^*}(p) = \left[(\boldsymbol{\mu} - r\mathbf{I})(\boldsymbol{\sigma}^2)^{-1} + 2\boldsymbol{\beta}\mathbf{K}(t)\right] \mathbf{Y}^{-1}(x - w) + \frac{\lambda(t)}{2} (\boldsymbol{\sigma}^2)^{-1} \left(\mathbf{H}(t) \exp(\mathbf{K}(t)\widetilde{\mathbf{Y}})\right) \mathbf{Y}^{-1} \hat{h}'(1 - p), \tag{3.27}$$

where $\mathbf{Q}_\Pi(p) = (\mathbf{Q}_{\Pi_1}(p), \cdots, \mathbf{Q}_{\Pi_n}(p))^\top$, $\boldsymbol{\mu} = \text{diag}\,(\mu_1, \cdots \mu_n)$, $\boldsymbol{\sigma}^2 = \text{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right)$, $\boldsymbol{\beta} = \text{diag}\,(\beta_1, \cdots \beta_n)$, $\mathbf{Y}^{-1} = \left(y_1^{-1}, \cdots, y_n^{-1}\right)^\top$. *In matrix form, the mean and standard of the optimal strategy under distinct market regimes can be expressed*

$$\mathbf{M}^* = \left(M_1^*, M_2^*, \cdots M_n^*\right)^\top = \left[(\boldsymbol{\mu} - r\mathbf{I})\left(\boldsymbol{\sigma}^2\right)^{-1} + 2\boldsymbol{\beta}\mathbf{K}(t)\right]\mathbf{Y}^{-1}(x - w), \tag{3.28}$$

$$\mathbf{N}^* = \left(N_1^*, N_2^*, \cdots N_n^*\right)^\top = \frac{\lambda(t)}{2}(\boldsymbol{\sigma}^2)^{-1}\exp\left(\int_t^T (\mathbf{Q} + \mathbf{C}\mathbf{K}(\tau))d\tau\right)\exp(\mathbf{K}(t)\widetilde{\mathbf{Y}})\mathbf{Y}^{-1}\|\hat{h}'\|_2, \tag{3.29}$$

$\mathbf{K}(t)$ *and* $\mathbf{H}(t)$ *are determined by equations* (3.16) (3.17). *Through the constraint* $E(X_T^{\Pi^*}) = z$, *the optimal Lagrange multiplier with start regime* e_i *at time* t *is given by* $w^* = \frac{z - x_0\exp\left\{\left(\frac{u_i - r}{\sigma_i^2} + 2\beta_i k_i\right)s^{2\beta_i}T\right\}}{1 - \exp\left\{\left(\frac{u_i - r}{\sigma_i^2} + 2\beta_i k_i\right)s^{2\beta_i}T\right\}}$.

Theorem 3.1 reveals that the strategy mean depends on the wealth level x, with only $\mathbf{K}(t)$ being time-dependent. Since $\mathbf{K}(t)$ increases over time, it follows that the mean exhibits a monotonically increasing trend with respect to time t. The standard deviation of the exploratory optimal strategy is independent of the wealth x and never diminishes to zero. Meanwhile the standard deviation declines with the volatility. Stochastic environments provide more learning opportunities, but because exploration is costly, the investor tends to reduce exploration. Notably, $\mathbf{M}^*$ does not depend on the exploration weight $\lambda(t)$, $\mathbf{N}^*$ is related to $\lambda(t)$. This result reveals a clear separation between exploitation and exploration. $\lambda(t)$ is a decreasing function of time. This means that the investor gradually focuses more on exploiting the existing strategy.

Theorem 3.2. *The optimal value function of the MV problem with regime switching of* (2.7) *is given by*

$$\mathbf{V}^{cl}(t, s, x) = \mathbf{a}(t, s)(x - w)^2 - (w - z)^2\mathbf{1}_n^\top, \tag{3.30}$$

where $\mathbf{V}^{cl}(t, s, x) = [V^{cl}(t, s, x, e_1), \cdots, V^{cl}(t, s, x, e_n)]^\top$, *and* $\mathbf{a}(t, s)$ *are determined by equations* (3.9). *The corresponding optimal strategy is given by* $\mathbf{u}_t^*$. *In matrix form, under distinct market regimes it can be expressed*

$$\mathbf{u}_t^* = (u^*(t, s, x, e_1), \cdots, u^*(t, s, x, e_n))^\top = \left[(\boldsymbol{\mu} - r\mathbf{I})\left(\boldsymbol{\sigma}^2\right)^{-1} + 2\boldsymbol{\beta}\mathbf{K}(t)\right]\mathbf{Y}^{-1}(x - w), \tag{3.31}$$

$K(t)$ *are determined by equations*(3.17). *The optimal Lagrange multiplier with start regime* e_i *at time* t *is given by* $w^* = \frac{z - x_0\exp\left\{\left(\frac{u_i - r}{\sigma_i^2} + 2\beta_i k_i\right)s^{2\beta_i}T\right\}}{1 - \exp\left\{\left(\frac{u_i - r}{\sigma_i^2} + 2\beta_i k_i\right)s^{2\beta_i}T\right\}}$.

When the exploration weight $\lambda(t)$ decreases to 0, the EMV problem with regime switching converges to the classical MV problem with regime switching. It follows directly that

$$\lim_{\lambda(t)\to 0}\Pi^*(\cdot; t, s, x, e_i) = u^*(\cdot; t, s, x, e_i),$$

and

$$\lim_{\lambda(t)\to 0}\left|V(t, s, x, e_i) - V^{cl}(t, s, x, e_i)\right| = 0.$$

According to Theorems 3.1 and 3.2, in the classical MV model, the investor focuses solely on exploitation-making optimal decisions based on currently known information without considering exploration of unknown environments. If a 'reasonably good' strategy is discovered early by chance, the investor may become trapped in a local optimum. As a result, they might consistently stick to this strategy and fail to discover the truly global optimal solution. In contrast, the exploratory mean-variance strategy under the reinforcement learning framework explicitly incorporates an exploration mechanism into the objective function through a regularizer. This approach not only encourages policy randomness to collect more environmental data but also continuously improves model quality to support better decision-making in the future, thereby achieving a deeper understanding of environmental characteristics through dynamic learning.

4. Policy iteration

Policy iteration is a dynamic programming algorithm used to solve Markov decision processes or stochastic control problems. It alternates between
(i) Policy evaluation: For any given policy, estimate its value function.
(ii) Policy improvement: Based on the value function of any given policy, update the policy.

Although Theorem 3.1 characterizes the optimal strategy distribution for the regime-switching EMV problem, its practical implementation relies on an iterative process. We typically begin with an initial policy guess, Π_0, and refine it successively until convergence. The policy improvement theorem (PIT) establishes an iterative update procedure that ensures a monotonic improvement or at least non-degradation of the policy performance.

Theorem 4.1. *[PIT] Let $w \in \mathbb{R}$ be fixed and $\Pi \in \mathcal{A}(\pi)$ be any given admissible feedback control strategy. Under the regularizer Φ_h, when the regime is e_i at time t the corresponding value function $V^{\Pi}(\cdot,\cdot,\cdot,e_i)$ satisfies $V^{\Pi}_{xx}(t,x,s,e_i) > 0$, $\forall (t,x,s,e_i) \in [0,T]\times\mathbb{R}\times\mathbb{R}\times\mathcal{N}$. We now construct a new strategy distribution $\widetilde{\Pi}$ whose regularizer is given by*

$$Q_{\widetilde{\Pi}}(p) = -\frac{(\mu_i - r)V^{\Pi}_x + \sigma_i^2 s^{2\beta_i+1} V^{\Pi}_{sx}}{\sigma_i^2 s^{2\beta_i} V^{\Pi}_{xx}} + \frac{\lambda(t)}{\sigma_i^2 s^{2\beta_i} V^{\Pi}_{xx}} \hat{h}'(1-p). \tag{4.1}$$

If the new strategy is feasible, then $\forall (t,x,s,e_i) \in [0,T]\times\mathbb{R}\times\mathbb{R}\times\mathcal{N}$, we have $V^{\widetilde{\Pi}}(t,x,s,e_i) \le V^{\Pi}(t,x,s,e_i)$.

See the Appendix for the proof.

Theorem 4.2. *For $p \in (0,1)$, let $\Pi^{(0)}$ denote the initial investment strategy, under which the regularizer is given by*

$$\mathbf{Q}^{(0)}_{\Pi}(p) = \left[(\boldsymbol{\mu} - r\mathbf{I})(\boldsymbol{\sigma}^2)^{-1} + 2\boldsymbol{\beta}\mathbf{K}^{(0)}(t)\right]\mathbf{Y}^{-1}(x-w) + \frac{\lambda(t)}{2}(\boldsymbol{\sigma}^2)^{-1}\left(\mathbf{H}^{(0)}(t)\exp(\mathbf{K}^{(0)}(t)\widetilde{\mathbf{Y}})\right)\mathbf{Y}^{-1}\hat{h}'(1-p). \tag{4.2}$$

For all $m \ge 1$, the strategy $\Pi^{(m)}(t,x,s)$ is updated according to the following equation

$$\mathbf{Q}^{(m)}_{\Pi}(p) = \left[(\boldsymbol{\mu} - r\mathbf{I})(\boldsymbol{\sigma}^2)^{-1} + 2\boldsymbol{\beta}\mathbf{K}^{(m)}(t)\right]\mathbf{Y}^{-1}(x-w) + \frac{\lambda(t)}{2}(\boldsymbol{\sigma}^2)^{-1}\left(\mathbf{H}^{(m)}(t)\exp(\mathbf{K}^{(m)}(t)\widetilde{\mathbf{Y}})\right)\mathbf{Y}^{-1}\hat{h}'(1-p),$$

where $(\mathbf{K}^{(m)}, \mathbf{H}^{(m)})$ *satisfy the following iterative formula*

$$\begin{cases} \dot{\mathbf{K}}^{(m)}(t) = -\mathbf{A} - \mathbf{B}\mathbf{K}^{(m)}(t) + \mathbf{E}\mathbf{K}^{(m-1)}(t)\mathbf{K}^{(m-1)}(t), & \mathbf{K}^{(m)}(T) = \mathbf{0}, \\ \dot{\mathbf{H}}^{(m)}(t) = -\mathbf{Q}\mathbf{H}^{(m)}(t) - \mathbf{C}\mathbf{K}^{(m-1)}(t)\mathbf{H}^{(m-1)}(t), & \mathbf{H}^{(m)}(T) = \mathbf{I}, \end{cases} \tag{4.3}$$

Then

$$\lim_{m\to\infty} \left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty = 0, \tag{4.4}$$

$$\lim_{m\to\infty} \left\|\mathbf{H}^{(m)}(t) - \mathbf{H}^*(t)\right\|_\infty = 0, \tag{4.5}$$

where

$$\left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty := \max\left\{\left|\left(k_i^{(m)}(t) - k_i^*(t)\right)\right|, 1 \le i \le n\right\},$$

$$\left\|\mathbf{H}^{(m)}(t) - \mathbf{H}^*(t)\right\|_\infty := \max\left\{\left|\left(h_i^{(m)}(t) - h_i^*(t)\right)\right|, 1 \le i \le n\right\}.$$

Proof. i) We first prove $\lim_{m\to\infty} \left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty = 0$. Recall $k_i(t)$ is (i,i)-th entry of the diagonal matrix $\mathbf{K}(t)$. It only needs to prove $\lim_{m\to\infty} |k_i^{(m)}(t) - k_i^*(t)| = 0$. Without loss of generality, let $T = 1$ and $W := \max_m |k_i^{(m)}(t)|$. Let $\delta_i^{(m)}(t) = k_i^{(m)}(t) - k_i^*(t)$. For $\delta_i^{(m+1)}(t)$, we have

$$\dot{\delta}_i^{(m+1)}(t) = \dot{k}_i^{(m+1)}(t) - \dot{k}_i^*(t) = 2\sigma_i^2\beta_i^2\left(k_i^{(m)}(t) - k_i^*(t)\right)\left(k_i^{(m)}(t) + k_i^*(t)\right) + 2(2r - \mu_i)\beta_i\delta_i^{(m+1)}(t).$$

where $\dot{\delta}_i^{(m+1)}(t)$ is the derivative of $\delta_i^{(m+1)}(t)$ with respect to t. Integrating both sides yields

$$\delta_i^{(m+1)}(t) = \int_t^1 e^{2(2r-\mu_i)\beta_i(s-t)}(-2\sigma_i^2\beta_i^2)\left(k_i^{(m)}(s) - k_i^*(s)\right)\left(\delta_i^{(m)}(s) + 2k_i^*(s)\right)ds.$$

Let $C_1 = 2\sigma_i^2\beta_i^2 e^{2|2r-\mu_i\|\beta_i|}$, thus

$$\begin{aligned} \left|\delta_i^{(m+1)}(t)\right| &\le 2\sigma_i^2\beta_i^2 \int_t^1 e^{2|2r-\mu_i\|\beta_i|}\left|\delta_i^{(m)}(s)\right|\left(2W + \left|\delta_i^{(m)}(s)\right|\right)ds \\ &= C_1 \int_t^1 \left|\delta_i^{(m)}(s)\right|\left(2W + \left|\delta_i^{(m)}(s)\right|\right)ds. \end{aligned}$$

Assume $\left|k_i^{(0)} - k_i^*\right| \le \overline{w}$. To prove that $\delta_i^{(m)}(t)$ converges to zero as $m \to \infty$, we introduce a sequence $\{L_m\}_m$ with $L_0 = \overline{w}$, which satisfies the recursive relation

$$L_{m+1} = \frac{\epsilon}{m+1}L_m + \frac{\zeta}{m+1}L_m^2,$$

with $\epsilon = 2\sigma_i^2\beta_i^2 e^{2|2r-\mu_i\|\beta_i|}2\left|k_i^*(t)\right|$ and $\zeta = 2\sigma_i^2\beta_i^2 e^{2|2r-\mu_i\|\beta_i|}$. We argue that $\left|\delta_i^{(m)}(t)\right| \le L_m(1-t)^m$, in fact it clearly holds when $m = 0$, if it is true for m, then we have

$$\begin{aligned} \left|\delta_i^{(m+1)}(t)\right| &\le C_1 \int_t^1 L_m(1-s)^m\left(2W + L_m^2(1-s)^{2m}\right)ds \\ &= C_1\frac{L_m(1-t)^{m+1}(2W)}{m+1} + C_1\frac{L_m^2(1-t)^{2m+1}}{2m+1} \\ &\le \left(\frac{C_1L_m2W}{m+1} + \frac{L_m^2C_1}{m+1}\right)(1-t)^{m+1} \\ &= L_{m+1}(1-t)^{m+1}. \end{aligned}$$

For the sequence $\{L_m\}_m$ with $L_0 = \overline{w}$, we see that

$$\frac{L_{m+1}}{L_m} = \frac{\epsilon + \zeta L_m}{m+1}, \quad L_{m+1} = \frac{\overline{w}}{(m+1)!}\prod_{i=0}^{m+1}(\epsilon + \zeta L_i).$$

Specially choose $\overline{w} < 1$ such that $\overline{w}\sup_m \frac{(\epsilon+\zeta)^m}{(m+1)!} < 1$, then $L_m \le 1, L_m \le \overline{w}\frac{(\epsilon+\zeta)^m}{m!}$ and

$$\lim_{m\to\infty}\left|\delta_i^{(m)}(t)\right| \le \lim_{m\to\infty} L_m(1-t)^m \le \lim_{m\to\infty}\overline{w}\frac{(\epsilon+\zeta)^m(1-t)^m}{m!} = 0.$$

This gives

$$\lim_{m\to\infty}\left|\delta_i^{(m)}(t)\right| = 0,$$

which implies

$$\lim_{m\to\infty}\left|k_i^{(m)}(t) - k_i^*(t)\right| = 0, \ \lim_{m\to\infty}\left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty = 0.$$

ii) We prove $\lim_{m\to\infty}\left\|\mathbf{H}^{(m)}(t) - \mathbf{H}^*(t)\right\|_\infty = 0$. Similarly, define $\boldsymbol{\xi}^{(m)}(t) := \mathbf{H}^{(m)}(t) - \mathbf{H}^*(t)$ and $\left\|\boldsymbol{\xi}^{(m)}(t)\right\|_\infty := \max\left\{\left|\left(\xi_i^{(m)}(t) - \xi_i^*(t)\right)\right|, 1 \le i \le n\right\}$. For $\dot{\boldsymbol{\xi}}^{(m+1)}(t)$, we have

$$\left\|\dot{\boldsymbol{\xi}}^{(m+1)}(t)\right\|_\infty = \left\|\dot{\mathbf{H}}^{(m+1)}(t) - \dot{\mathbf{H}}^*(t)\right\|_\infty = \left\|-\mathbf{Q}\boldsymbol{\xi}^{(m+1)}(t) - \mathbf{C}(\mathbf{K}^{(m)}(t)\boldsymbol{\xi}^{(m)}(t) - \mathbf{C}\left(\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right)\mathbf{H}^*(t))\right\|_\infty,$$

taking the integral, we have

$$\begin{aligned}
\left\|\boldsymbol{\xi}^{(m+1)}(t)\right\|_\infty &= \left\|\int_t^1 e^{\mathbf{Q}(s-t)}\left[\mathbf{C}\mathbf{K}^{(m)}(s)\boldsymbol{\xi}^{(m)}(s) + \mathbf{C}\left(\mathbf{K}^{(m)}(s) - \mathbf{K}^*(s)\right)\mathbf{H}^*(s)\right]ds\right\|_\infty \\
&\le \int_t^1 \left\|e^{\mathbf{Q}(s-t)}\right\|_\infty \|\mathbf{C}\|_\infty\left[\left\|\mathbf{K}^{(m)}(s)\right\|_\infty\left\|\boldsymbol{\xi}^{(m)}(s)\right\|_\infty + \left\|\left(\mathbf{K}^{(m)}(s) - \mathbf{K}^*(s)\right)\right\|_\infty\left\|\mathbf{H}^*(s)\right\|_\infty\right]ds \\
&\le \max_{t\le s\le 1}\left\|e^{\mathbf{Q}(1-t)}\right\|_\infty \|\mathbf{C}\|_\infty\int_t^1\left[\left\|\mathbf{K}^{(m)}(s)\right\|_\infty\left\|\boldsymbol{\xi}^{(m)}(s)\right\|_\infty + \left\|\left(\mathbf{K}^{(m)}(s) - \mathbf{K}^*(s)\right)\right\|_\infty\left\|\mathbf{H}^*(s)\right\|_\infty\right]ds.
\end{aligned}$$

By Gronwall's inequality we get

$$\left\|\boldsymbol{\xi}^{(m+1)}(t)\right\| \le e^{\left\|e^{\mathbf{Q}(1-t)}\right\|_\infty\|\mathbf{C}\|_\infty\int_t^1\left\|\mathbf{K}^{(m)}(s)\right\|_\infty ds}\max_{t\le s\le 1}\left\|e^{\mathbf{Q}(1-t)}\right\|_\infty\|\mathbf{C}\|_\infty\int_t^1\left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty\left\|\mathbf{H}^*(s)\right\|_\infty ds.$$

The earlier result $\lim_{m\to\infty}\left\|\mathbf{K}^{(m)}(t) - \mathbf{K}^*(t)\right\|_\infty = \mathbf{0}$ and $\left\|\mathbf{K}^{(m)}(t)\right\|_\infty$ is bounded together to give $\lim_{m\to\infty}\|\boldsymbol{\xi}^{(m+1)}(t)\|_\infty = \mathbf{0}$, thus

$$\lim_{m\to\infty}\|\mathbf{H}^{(m)}(t) - \mathbf{H}^*(t)\|_\infty = 0.$$

5. Algorithm design

We develop an RL algorithm to solve the EMV problem with regime switching. The investor knows the risk-free rate r and the exploration weight $\lambda(t)$. However, the investor may not be able to accurately evaluate the volatilities and the returns, and might not even know which type of CEV model the risky asset follows. All he can rely on are the historical observations of (S_t, α_t).

In numerical experiments, we consider investment strategies following a normal distribution. Let $\hat{h}(p) = \int_0^p z(1-s)ds$, where z is the standard normal distribution function. According to Han et al. [25], we have $\Phi_{\hat{h}}(\Pi) = \int_0^1 Q_\Pi(p)z(p)dp$. In this case, $\|\hat{h}'\|_2^2 = 1$. The mean and standard deviation corresponding to the optimal investment strategy under n regimes are given by

$$\mathbf{M}^* = \left(M_1^*, M_2^*, \cdots M_n^*\right)^\top = \left[(\boldsymbol{\mu} - r\mathbf{I})\left(\boldsymbol{\sigma}^2\right)^{-1} + 2\boldsymbol{\beta}\mathbf{K}(t)\right]\mathbf{Y}^{-1}(x-w), \tag{5.1}$$

$$\mathbf{N}^* = \left(N_1^*, N_2^*, \cdots N_n^*\right)^\top = \frac{\lambda(t)}{2}(\boldsymbol{\sigma}^2)^{-1}\exp\left(\int_t^T (\mathbf{Q}+\mathbf{C}\mathbf{K}(\tau))d\tau\right)\exp(\mathbf{K}(t)\widetilde{\mathbf{Y}})\mathbf{Y}^{-1}, \tag{5.2}$$

For numerical illustration, we consider a financial market with only two regimes $\mathcal{N} = \{e_1, e_2\}$, and assume $\beta_1 = 0$, $\beta_2 = -\frac{1}{2}$, which means the risky asset price switches between the GBM and the CIR process. To implement the algorithm, we employ a discrete-time approximation of the problem under study. First, we discretize the continuous time horizon $[0,T]$ into N equal-length intervals, $\Delta t = t_{k+1} - t_k$, $k = 0, 1, \ldots, N-1$. The investor has collected the historical data of the discrete observations. Based on the historical information, he follows the strategy $\Pi(t_k)$ at time t_k and draws an action $u(t_k)$ from $\Pi(t_k)$. The discounted wealth process at the discretized time t_{k+1} is given by

$$X^\pi(t_{k+1}) \approx X^\pi(t_k) + u(t_k)\frac{e^{-rt_{k+1}}S(t_{k+1}) - e^{-rt_k}S(t_k)}{e^{-rt_k}S(t_k)}. \tag{5.3}$$

We can obtain a series of samples $D = \{(t_\tau, x_\tau, s_\tau, \alpha_\tau), \tau = 0, 1, \ldots, N-1\}$. In reinforcement learning, common methods for value function and policy parameterization include linear function approximation (Sutton (1988) [35]) and nonlinear function approximation using neural networks (Mnih et al. (2015) [36]). We adopt the linear function approximation approach to parameterize the value function and the policy. If the initial regime is e_i, the following expression is used to approximate the strategy defined in (3.28)–(3.29) as well as the value function in (3.26).

$$\begin{aligned} V^\Theta(t,x,s,e_i) = {} & p(\theta_i^{(V,0)}, T-t)e^{p(\theta_i^{(V,1)}, T-t)s^{-2\beta_i}}(x-w)^2 + p(\theta_i^{(V,2)}, T-t)s^{-4\beta_i} \\ & + p(\theta_i^{(V,3)}, T-t)s^{-2\beta_i} + p(\theta_i^{(V,4)}, T-t), \end{aligned} \tag{5.4}$$

$$M_i^* = -\frac{\phi V_x^\Theta + \xi s^{2\beta_i+1}V_{sx}^\Theta}{\xi s^{2\beta_i}V_{xx}^\Theta}, \quad N_i^* = \frac{\lambda(t)\left\|\hat{h}'\right\|_2}{\xi s^{2\beta_i}V_{xx}^\Theta}. \tag{5.5}$$

The policy parameters are $\phi = (\mu_i - r)$ and $\xi = \sigma_i^2$. The function $p(\theta, t)$, where $\theta \in \mathbb{R}^d$ is a parameter vector, typically consists of the first d terms of a Taylor series expansion. Let $\Theta = (\theta_i^{(V,0)}, \theta_i^{(V,1)}, \theta_i^{(V,2)}, \theta_i^{(V,3)}, \theta_i^{(V,4)})$. The components of Θ represent the corresponding Taylor coefficients. In this paper, we update these parameters using the temporal difference (TD) learning which was initially introduced by Sutton (1988) [35]. From Bellman's optimality principle, we have

$$V^\pi(t,x,s,e_i) = E\left[V^\pi(h, X_h, S_h, \alpha_h) - \int_t^h \lambda(v)\Phi(v)dv \middle| X_t = x, S_t = s, \alpha_t = e_i\right].$$

For fixed $(t,x,s,e_i) \in [0,T]\times\mathbb{R}\times\mathbb{R}\times\mathcal{N}$, dividing both sides by $h-t$ and rearranging terms yields

$$E\left[\frac{V^\pi(h, X_h, S_h, \alpha_h) - V^\pi(t,x,s,e_i)}{h-t} - \frac{1}{h-t}\int_t^h \lambda(v)\Phi(v)dv \middle| X_t = x, S_t = s, \alpha_t = e_i\right] = 0.$$

When $h = t + \Delta t$, the Bellman error is defined as

$$\delta_t := \dot{V}_t^{\Pi} - \lambda(t)\Phi(\Pi),$$

where $\dot{V}_t^{\pi} = \frac{V^{\pi}(t+\Delta t, X_{t+\Delta t}, S_{t+\Delta t}, \alpha_{t+\Delta t}) - V^{\pi}(t,x,s,e_i)}{\Delta t}$, with Δt being the discretization step size of the learning algorithm. The TD loss is defined as the mean squared TD error evaluated at discrete time points $t_{k=0,\dots,K-1}$. Via the parameterized value function and policy, it can be expressed as

$$TD(\Theta,\phi,\xi) = \frac{1}{2}E\left[\sum_{k=0}^{K-1}\left(\frac{V^{\Theta}(t+\Delta t, X_{t+\Delta t}, S_{t+\Delta t}, \alpha_{t+\Delta t}) - V^{\Theta}(t,x,s,e_i)}{\Delta t} - \lambda(t)\Phi^{(\Theta,\phi,\xi)}(\Pi)\right)^2 \Delta t\right]. \tag{5.6}$$

The market parameters Θ, ϕ, ξ are updated via gradient descent

$$\Theta^{(n+1)} \leftarrow \Theta^{(n)} - \eta_{\Theta}\nabla_{\Theta} TD(\Theta,\phi,\xi),$$
$$\phi^{(n+1)} \leftarrow \phi^{(n)} - \eta_{\phi}\nabla_{\phi} TD(\Theta,\phi,\xi),$$
$$\xi^{(n+1)} \leftarrow \xi^{(n)} - \eta_{\xi}\nabla_{\xi} TD(\Theta,\phi,\xi).$$

Finally, the Lagrange multiplier is updated according to the terminal condition $E[X_T] = z$

$$w_{n+1} = w_n - \eta_w(X_T - z),$$

where the learning rate satisfies $\eta_w > 0$.

algorithm 1 Exploratory mean-variance portfolio under regime switching

Input: Learning rates η_{Θ}, η_{ϕ}, η_{ξ}, η_w, sample average size N, investment horizon T, time step size $\Delta t = T/N$, exploration weight $\lambda(t)$, initial wealth x_0, iteration count for updating the Lagrange multiplier w, number of iterations M.
Initialize Θ, ϕ, ξ and w.
for $k = 1 \to M$ **do**
  for $i = 1 \to \left\lfloor \frac{T}{\Delta t} \right\rfloor$ **do**
    The tuple $\left(t_i^k, X_i^k, S_i^k, \alpha_i^k\right)$ is sampled from the policy $\Pi^{(\phi,\xi)}$.
    Acquire all data sets $\mathcal{D} = \left\{\left(t_i^k, X_i^k, S_i^k, \alpha_i^k\right), 1 \le i \le \left\lfloor \frac{T}{\Delta t} \right\rfloor\right\}$.
    Compute V^{Θ}.
    Update

$$\Theta^{(k+1)} \leftarrow \Theta^{(k)} - \eta_{\Theta}\nabla_{\Theta} TD(\Theta,\phi,\xi),$$
$$\phi^{(k+1)} \leftarrow \phi^{(k)} - \eta_{\phi}\nabla_{\phi} TD(\Theta,\phi,\xi),$$
$$\xi^{(k+1)} \leftarrow \xi^{(k)} - \eta_{\xi}\nabla_{\xi} TD(\Theta,\phi,\xi).$$

  end for
  Update $\Pi^{(\phi,\xi)}$ from (5.5).
  if $k \bmod N == 0$ **then**
    Update $w \leftarrow w - \eta_w\left(\frac{1}{N}\sum_{j=k-N+1}^{k} X^j_{\lfloor \frac{T}{\Delta t} \rfloor} - z\right)$.
  end if
end for

Numerical experiments based on simulated data are conducted to validate the proposed algorithm. The regularizer is selected such that the randomized policy follows a normal distribution. The investment horizon is fixed at to $T = 1$ with time step $\Delta t = \frac{1}{252}$ (corresponding to $N = 252$ steps for the annual MV problem). The annualized interest rate is set to $r = 0.02$. For both the GBM and the CIR process, the annualized return μ takes values in $\{0.03, 0.05\}$ and volatility σ in $\{0.1, 0.2\}$. The initial wealth is $X_0 = 1$, with a target annualized terminal return of 40% ($z = 1.4$). Assume the Markov regime transition intensity rate matrix as $\mathbf{Q} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$. Parametrize (5.4) and (5.5) by selecting $p(\theta, T-t)$ as

$$p(\theta, t) = \theta_0 t^2 + \theta_1 t + \theta_2.$$

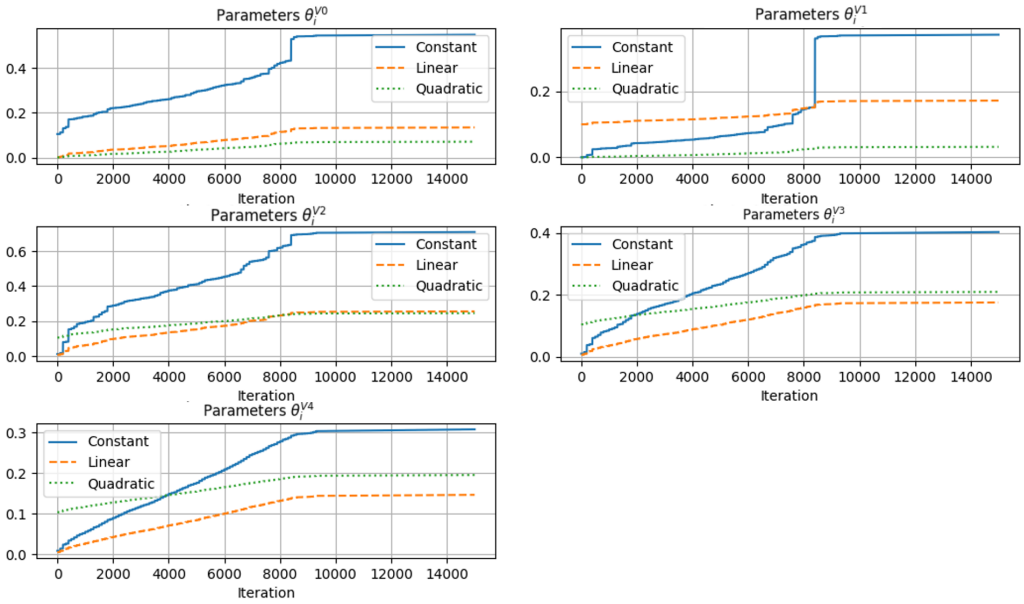

Figure 1. Parameters of the Taylor expansion in the value function.

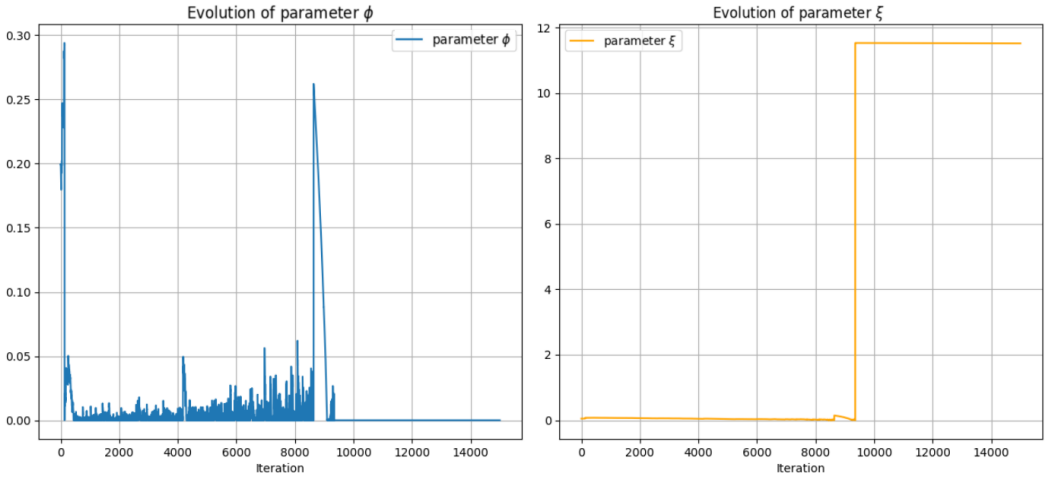

Figure 2. The parameters in the Policy.

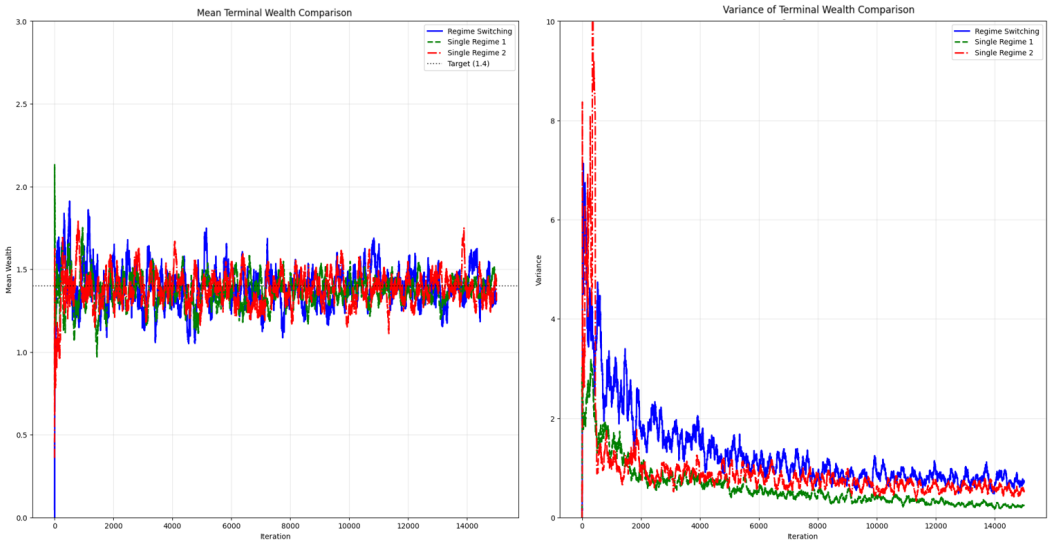

Figure 3. The mean and variance of exploratory strategies.

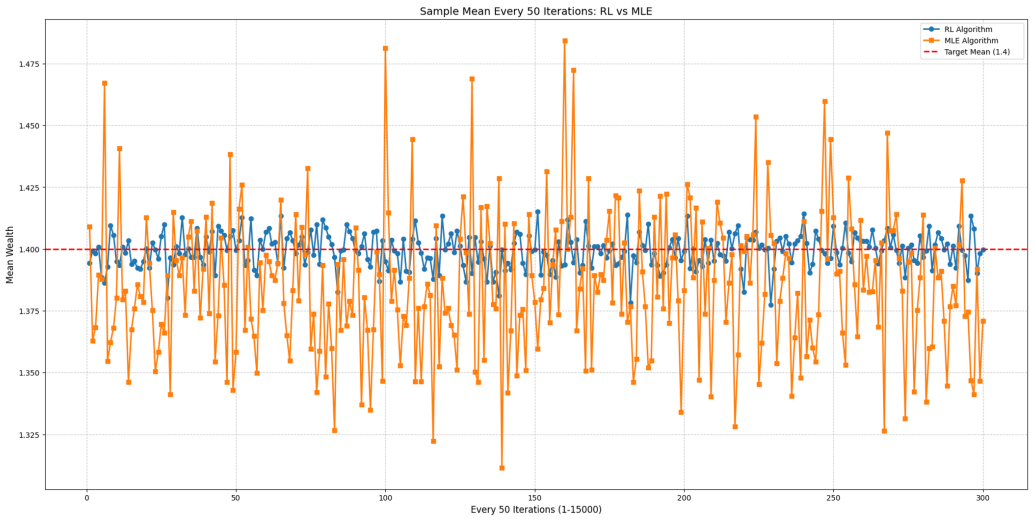

Figure 4. Learning curves of the sample mean of terminal wealth.

From Figures 1 and 2, we can observe that, after 8500 iterations, the Taylor expansion parameters of the value function stabilize, showing good convergence. The policy parameters ϕ and ξ stabilize after 9000 iterations. At this point, both the policy and value function exhibit stable convergence behavior. Figure 3 compares the mean-variance performance of the optimal strategy under the risky asset models of GBM, CIR, and regime switching between the two models. Figure 3 shows that after 1000 iterations, the mean of the exploration strategy converges to the target value of 1.4. Compared to the single-regime market, the regime-switching strategy exhibits larger fluctuations in the mean, indicating lower stability. The variance of the strategy converges after approximately 12,000 iterations. Initially, the variance of the CIR process is larger, indicating a broader exploration range. However, as the iterations progress, the variance rapidly decreases and becomes smaller than that of the regime-switching strategy, with the GBM variance consistently remaining low. This may be due to

the regime-switching market intensifying exploration, leading to higher variance. Additionally, the higher exploration cost associated with regime switching slows down the convergence of the variance. Figures 4 and 5 compare the sample means and variances of terminal wealth, calculated every 50 iterations, between the RL algorithm proposed in this paper and the traditional MLE method. The results show that the RL algorithm exhibits more stable convergence, with smaller variance in terminal wealth concentrated around the target value of 1.4, demonstrating superior learning performance.

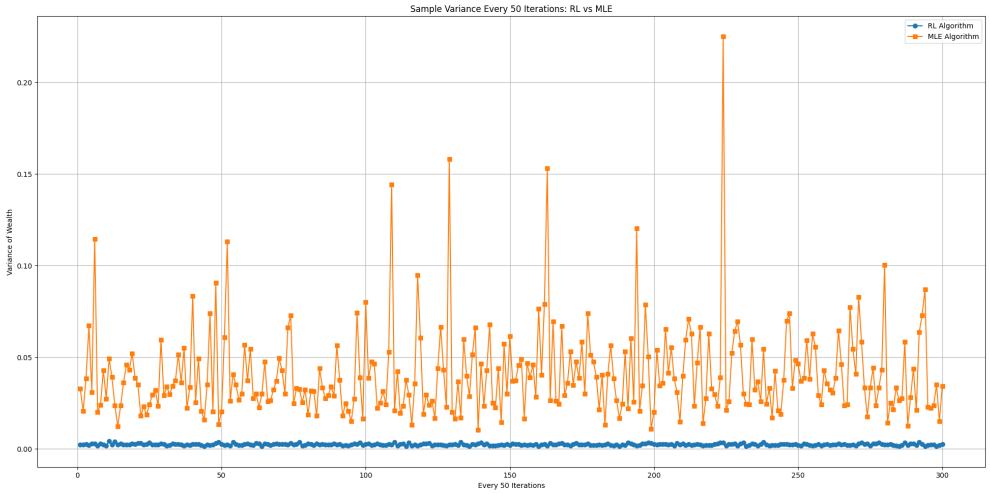

Figure 5. Learning curves of the sample variance of terminal wealth.

6. Conclusion

We introduce a continuous-time reinforcement learning framework for a financial market with Markov regime switching. In this setting, the risky asset transitions among n distinct CEV models, with market coefficients evolving according to the prevailing regime. The regime dynamics are governed by a continuous-time Markov chain. Within the reinforcement learning framework, we establish an EMV problem for the investor, and derive the optimal strategy together with the corresponding value function across n different CEV models via the dynamic programming approach. To identify the optimal strategy within a reinforcement learning framework, we construct a policy iteration scheme and prove the convergence of the policy. From an algorithmic and implementation perspective, we concentrate on a setting in which the model switches between a GBM and a CIR process, with a given regularizer that ensures the policy follows a normal distribution. The policy and value function are parameterized using linear function approximation, and numerical experiments are conducted to demonstrate the convergence of the parameters.

Author contributions

Xiaoyu Xing: Methodology, Supervision; Xingtian Zhang: Writing-original draft, Formal analysis, Validation, Visualization, Conceptualization, Writing-review and editing; Jiarou Luo: Writing-review and editing, Validation.

Use of Generative-AI tools declaration

The authors declare they have used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 12071107).

Conflict of interest

The authors declare that they have no conflicts of interest relevant to this study.

References

1. H. Markowitz, Portfolio selection, *J. Finance,* **7** (1952), 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x
2. D. Li, W. L. Ng, Optimal dynamic portfolio selection: multiperiod mean-variance formulation, *Math. Finance,* **10** (2000), 387–406. https://doi.org/10.1111/1467-9965.00100
3. X. Y. Zhou, D. Li, Continuous-time mean-variance portfolio selection: a stochastic LQ framework, *Appl. Math. Optim.,* **42** (2000), 19–33. https://doi.org/10.1007/s002450010003
4. M. C. Chiu, D. Li, Asset and liability management under a continuous-time mean-variance optimization framework, *Insur. Math. Econ.,* **39** (2006), 330–355. https://doi.org/10.1016/j.insmatheco.2006.03.006
5. J. N. Zhang, P. Chen, Z. Jin, S. Li, Open-loop equilibrium strategy for mean-variance portfolio selection: a log-return model, *J. Ind. Manag. Optim.,* **17** (2021), 765–777. https://doi.org/10.3934/jimo.2019133
6. J. D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica,* **57** (1989), 357–384. https://doi.org/10.2307/1912559
7. X. Y. Zhou, G. Yin, Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model, *SIAM J. Control Optim.,* **42** (2003), 1466–1482. https://doi.org/10.1137/S0363012902405583
8. V. A. Gal'perin, V. V. Dombrovsky, E. N. Fedosov, Dynamic control of the investment portfolio in the jump-diffusion financial market with regime switching, *Autom. Remote Control,* **66** (2005), 837–850. https://doi.org/10.1007/s10513-005-0127-9
9. P. Chen, H. L. Yang, G. Yin, Markowitz's mean-variance asset-liability management with regime switching: a continuous-time model, *Insur. Math. Econ.,* **43** (2008), 456–465. https://doi.org/10.1016/j.insmatheco.2008.09.001
10. S. X. Xie, Continuous-time mean-variance portfolio selection with liability and regime switching, *Insur. Math. Econ.,* **45** (2009), 148–155. https://doi.org/10.1016/j.insmatheco.2009.05.005
11. P. Chen, H. L. Yang, Markowitz's mean-variance asset-liability management with regime switching: a multi-period model, *Appl. Math. Finance,* **18** (2011), 29–50. https://doi.org/10.1080/13504861003703633

12. X. W. Chen, F. Z. Huang, X. F. Li, Robust asset-liability management under CRRA utility criterion with regime switching: a continuous-time model, *Stoch. Model,* **38** (2022), 167–189. https://doi.org/10.1080/15326349.2021.1985520

13. J. Zhou, G. Liu, H. L. Yang, Optimal consumption and investment strategies with liquidity risk and lifetime uncertainty for Markov regime-switching jump diffusion models, *Eur. J. Oper. Res.,* **280** (2020), 1130–1143. https://doi.org/10.1016/j.ejor.2019.07.066

14. J. Eisenberg, L. Fabrykowski, M. D. Schmeck, Optimal surplus-dependent reinsurance under regime-switching in a brownian risk model, *Risks,* **9** (2021), 73. https://doi.org/10.3390/risks9040073

15. D. X. Zhao, C. N. Song, Liability assessment of life insurance companies in regime switching market, *Commun. Stat-theor. M.,* **54** (2025), 7210–7229. https://doi.org/10.1080/03610926.2025.2467204

16. H. R. Wang, T. Zariphopoulou, X. Y. Zhou, Reinforcement learning in continuous time and space: a stochastic control approach, *J. Mach. Learn. Res.,* **21** (2020), 8145–8178. http://jmlr.org/papers/v21/19-144.html

17. H. Wang, X. Y. Zhou, Continuous-time mean-variance portfolio selection: A reinforcement learning framework, *Math. Finance,* **30** (2020), 1273–1308. https://doi.org/10.1111/mafi.12281

18. X. Guo, R. Y. Xu, T. Zariphopoulou, Entropy regularization for mean field games with learning, *Math. Oper. Res.,* **47** (2022), 3239–3260. https://doi.org/10.1287/moor.2021.1238

19. D. Firoozi, S. Jaimungal, Exploratory LQG mean field games with entropy regularization, *Automatica,* **139** (2022), 110177. https://doi.org/10.1016/j.automatica.2022.110177

20. Y. W. Jia, X. Y. Zhou, Policy evaluation and temporal-difference learning in continuous time and space: a martingale approach, *J. Mach. Learn. Res.,* **23** (2022), 1–55. http://jmlr.org/papers/v23/21-0947.html

21. Y. W. Jia, X. Y. Zhou, Policy gradient and actor-critic learning in continuous time and space: theory and algorithms, *J. Mach. Learn. Res.,* **23** (2022), 1–50. http://jmlr.org/papers/v23/21-1387.html

22. Y. W. Jia, X. Y. Zhou, Q-learning in continuous time, *J. Mach. Learn. Res.,* **24** (2023), 1–61. ahttp://jmlr.org/papers/v24/22-0755.html

23. M. Dai, Y. C. Dong, Y. W. Jia, Learning equilibrium mean-variance strategy, *Math. Financ.,* **33** (2023), 1166–1212. https://doi.org/10.1111/mafi.12402

24. M. Dai, Y. C. Dong, Y. W. Jia, X. Y. Zhou, Data-driven Merton's strategies via policy randomization, *arXiv preprint, arxiv: 2312.11797*, 2025. https://doi.org/10.48550/arXiv.2312.11797

25. X. Han, R. D. Wang, X. Y. Zhou, Choquet regularization for continuous-time reinforcement learning, *SIAM J. Contral Optim.,* **61** (2023), 2777–2801. https://doi.org/10.1137/22M1524734

26. J. Y. Guo, X. Han, H. Wang, Exploratory mean-variance portfolio selection with Choquet regularizers, *arXiv preprint, arXiv: 2307.03026*, 2023. https://doi.org/10.48550/arXiv.2307.03026

27. J. Guo, X. Han, H. Wang, K. C. Yuen, A non-zero-sum game with reinforcement learning under mean-variance framework, *arXiv preprint, arXiv: 2502.04788*, 2025. https://doi.org/10.48550/arXiv.2502.04788

28. Y. M. Chen, B. Li, D. Saunders, Exploratory mean-variance portfolio optimization with regime-switching market dynamics, *arXiv preprint, arXiv: 2501.16659*, 2025. https://doi.org/10.48550/arXiv.2501.16659

29. B. Wu, L. F. Li, Reinforcement learning for continuous-time mean-variance portfolio selection in a regime-switching market, *J. Econ. Dyn. Contral.,* **158** (2024), 104787. https://doi.org/10.1016/j.jedc.2023.104787

30. X. F. Li, D. X. Zhao, X. W. Chen, Asset-liability management with state-dependent utility in the regime-switching market, *Stoch. Models,* **39** (2023), 566–591. https://doi.org/10.1080/15326349.2022.2138440

31. S. Basak, G. Chabakauri, Dynamic mean-variance asset allocation, *Rev. Financ. Stud.,* **23** (2010), 2970–3016. https://doi.org/10.1093/rfs/hhq028

32. T. Björk, A. Murgoci, A general theory of Markovian time inconsistent stochastic control problems, *SSRN Elect. J.*, 2010. https://doi.org/10.2139/ssrn.1694759

33. R. D. Wang, Y. R. Wei, G. E. Willmot, Characterization, robustness and aggregation of signed Choquet integrals, *Math. Oper. Res.,* **45** (2020), 993–1015. https://doi.org/10.1287/moor.2019.1020

34. M. Jeanblanc, M. Yor, M. Chesney, *Mathematical methods for financial markets*, London: Springer, 2009. https://doi.org/10.1007/978-1-84628-737-4

35. R. S. Sutton, Learning to predict by the methods of temporal differences, *Mach. Learn.,* **3** (1988), 9–44. https://doi.org/10.1023/A:1022633531479

36. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, et al. Human-level control through deep reinforcement learning, *Nature,* **518** (2015), 529–533. https://doi.org/10.1038/nature14236

AIMS Press