arXiv:2611.01046v1 [cs.LG] 31 Oct 2026

ENTROPY-REGULARIZED REINFORCEMENT LEARNING FOR THE REGIME-SWITCHING CONSUMPTION-INVESTMENT PROBLEM*

CONGDE HU†, DANPING LI‡, XUN LI§, AND YIJIE ZENG§

Abstract. This paper establishes an entropy-regularized reinforcement learning (ERRL) framework for the continuous-time consumption-investment problem under a regime-switching market model with general utility functions. By incorporating randomized exploratory policies and Shannon entropy regularization into the stochastic optimal control formulation, we derive closed-form representations for the optimal exploratory consumption and investment strategies. Specifically, the optimal investment strategy follows a multivariate Gaussian distribution, whereas the optimal consumption strategy takes the form of a Gibbs measure, which reduces to a Gamma distribution under logarithmic utility. We also briefly illustrate the extension to power utility. Our analysis shows that the modes of the exploratory distributions recover the functional forms of the classical deterministic optimal controls, rather than generally coinciding with their means. Furthermore, under logarithmic utility, we establish the convergence of the exploratory value function to its classical unregularized counterpart as the temperature parameter vanishes. For algorithmic design, we develop a continuous-time q-learning framework grounded in a continuous-time q-function, prove the policy improvement theorem, and establish martingale characterizations for both on-policy and off-policy learning. We find that parameterizing the actor network based on theoretical distribution families can help accelerate convergence and enhance numerical stability. Furthermore, numerical experiments across regime-switching environments provide empirical support for our theoretical findings, demonstrating stable loss convergence as well as the monotonicity and concavity of the learned value functions.

Key words. Stochastic control, Reinforcement learning, Consumption-investment, Regime-switching, General utility, q-learning

MSC codes. 68Q25, 68R10, 68U05

1. Introduction.

The continuous-time consumption-investment problem pioneered by [10] provides a classical foundation for dynamic asset allocation under stochastic market environments. In Merton's canonical model, investors maximize the lifetime expected utility derived from intermediate consumption and terminal wealth by solving a Hamilton–Jacobi–Bellman (HJB) equation, which admits closed-form deterministic optimal consumption and portfolio rules under logarithmic or constant relative risk aversion (CRRA) utilities. Recent work can be found in [1], [8], [20] and so on. In the financial market, investors usually face two market regimes, One represents a bull market characterized by rising asset prices, whereas the other represents a bear market characterized by declining asset prices. Therefore, the portfolio selection problem with regime-switching is of great interest and importance in financial investment. Under regime-switching models, optimal strategies for utility maximization have been extensively studied under various preferences, such as [14], [18] and so on.

A practical limitation shared by classical stochastic optimal control approaches to consumption-investment problems is their full-information prerequisite: agents are assumed to perfectly observe and precisely estimate all market primitives, including Markov chain transition intensities, regime-dependent drifts and volatility matrices

*Submitted to the editors DATE.

Funding: This work was funded by XXXXXX.

†School of Statistics, East China Normal University, Shanghai, China (hcd.math@gmail.com, dpli@fem.ecnu.edu.cn).

‡Corresponding author.

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, China (malixun@polyu.edu.hk, yi-jie.zeng@connect.polyu.hk).

[11]. In empirical applications, these structural parameters are often latent or subject to non-negligible estimation errors, which may render deterministic strategies less reliable or difficult to implement for real-time investment decisions. This observation motivates the integration of reinforcement learning (RL) with continuous-time stochastic control, which enables model-free policy learning through repeated environment interactions without requiring prior full knowledge of market dynamics.

Early continuous-time RL theories rooted in stochastic control were developed by [16, 17], who introduced entropy-regularized exploratory policies for mean-variance portfolio selection and established a link between soft RL and HJB variational problems. Subsequent work provided martingale characterizations for temporal difference (TD) evaluation and continuous-time q-learning, offering a theoretical basis for actor-critic algorithms on state spaces [5, 6, 19]. Extensions to alternative entropy regularizers have further illustrated the potential of exploratory stochastic control for financial tasks [7, 15]. Despite these advances, the existing continuous-time RL literature has paid limited attention to the joint optimization of consumption and investment strategies within regime-switching markets, and to the statistical structure of optimal exploratory consumption and investment distributions under utility functions.

This paper develops an entropy-regularized reinforcement learning (ERRL) framework for the continuous-time consumption-investment optimization problem with general concave utility functions in a regime-switching diffusion market. We embed Shannon entropy regularization into the classical stochastic control objective to induce randomized exploratory policies, which relaxes the complete market information requirement of traditional Merton-type models. Our contributions are organized along theoretical, structural, and algorithmic-numerical dimensions.

First, we reformulate the regime-switching consumption-investment problem as an entropy-regularized stochastic optimal control problem with stochastic feedback policies, and derive the associated variational HJB equation for the regularized value function. Via calculus of variations over probability measures, the joint optimal exploratory action distribution can be separated into independent portfolio weight and consumption marginals: the optimal risky asset allocation follows a multivariate Gaussian law, while consumption obeys a Gibbs measure. An important contribution from our derivation reveals that the classical deterministic Merton optimal controls correspond to the *mode* of the corresponding probability density, rather than the mean, of these exploratory distributions—a distinction that is particularly relevant for skewed consumption distributions such as the Gamma law under logarithmic utility. Under logarithmic utility, we also establish uniform convergence of the entropy-regularized value function to its classical unregularized counterpart as the temperature parameter vanishes, thereby connecting the exploratory RL framework with standard utility maximization theory.

Second, we investigate the optimal exploratory consumption policy under logarithmic utility and show that it follows a Gamma distribution with explicit shape and rate parameters. We further establish the convergence of the exploratory value function to its classical counterpart as the temperature parameter vanishes. As an additional illustration, we briefly discuss the corresponding Gibbs density under power utility and identify its mode. These results provide theoretical guidance for selecting suitable policy distributions in reinforcement learning algorithms.

Third, we construct and numerically validate a model-free continuous-time q-learning actor-critic algorithm adapted to regime-switching markets. Algorithmically, we introduce a regime-adjusted continuous-time q-function, establish a policy im-

provement result under entropy regularization, and derive martingale representations for both on-policy and off-policy learning pipelines. The proposed actor network employs a Gaussian distribution for investment and a Gamma-based parameterization for consumption, providing a tractable implementation of exploratory policies. To illustrate the applicability of our framework, we conduct numerical experiments on a two-state bull–bear simulator using logarithmic and power utilities. The results suggest stable training convergence and indicate that the learned value functions are generally consistent with economic properties, such as risk-averse concavity and a clear regime hierarchy (with bull-market values generally lying above bear-market values). Furthermore, the learned exploratory policies exhibit patterns consistent with the theoretical policy structures, reflecting intuitive consumption and investment shifts across alternating market conditions.

The structure of this paper is organized as follows. Section 2 constructs the theoretical framework for the exploratory consumption-investment problem under the regime-switching model and derives the optimal exploratory policy. Section 3 investigates the logarithmic utility case and briefly discusses its extension to power utility. Section 4 designs a q-learning algorithm and provides numerical examples to illustrate our results. Finally, Section 5 concludes the paper.

Notation: Throughout the paper, by convention, all vectors are column vectors unless otherwise specified. Let $\mathbb{R}^N$ be the N-dimensional Euclidean space and $\mathbb{R}^{N\times N}$ be the space of real $N \times N$ matrices. For a matrix A, $A^\top$ is its transpose, A^{-1} its inverse, $\det(A)$ its determinant, and $\|A\|_F$ its Frobenius norm. For a vector x, $\|x\|_2$ denotes its Euclidean norm. We use I to denote the identity matrix and $\mathbf{1}$ to denote a vector of ones with appropriate dimensions. For a sufficiently smooth function $v(t,x,i)$, v_t, v_x, and v_{xx} denote the first-order partial derivative with respect to time t, and the first- and second-order partial derivatives with respect to the spatial variable x, respectively. The symbol $\mathbb{E}_{t,x,i}[\cdot]$ denotes the conditional expectation given $X(t) = x$, $\alpha(t) = i$. The symbol δ is used to denote the variational operator. Let $\mathcal{A}$ denote the action space, and $\mathcal{P}(\mathcal{A})$ denote the set of probability density functions defined on $\mathcal{A}$. Finally, $\mathcal{N}(\cdot \mid \mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, and $\text{Gamma}(\cdot \mid \alpha, \beta)$ denotes a Gamma distribution with shape parameter α and rate parameter β.

2. Model formulation.

This section presents a mathematical formulation for the stochastic optimal control problem of exploratory consumption-investment decision. To provide a baseline for comparison, we first review the classical consumption-investment problem, following the setup in [4], and then introduce the exploratory consumption-investment problem by incorporating policy exploration into the decision process. Finally, we discuss the theoretical characterization of the proposed problem and derive the explicit form of the optimal policy.

2.1. Classical consumption-investment problem.

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete probability space. We consider a continuous-time Markov chain $\alpha(t)$ taking values in a finite state space $\mathcal{M} = \{1, 2, \ldots, D\}$. The generator matrix of $\alpha(t)$ is given by $Q = (q_{ij})_{D\times D}$, such that for $i \neq j$, $q_{ij} \geq 0$ represents the transition rate from state i to state j, and $q_{ii} = -\sum_{j\neq i} q_{ij}$. The market consists of one risk-free asset (bank account) and N risky assets (stocks) traded continuously over a finite time horizon $[0, T]$. The market parameters depend on the current regime state $i \in \mathcal{M}$. For state $i \in \mathcal{M}$, let $r(t,i)$ be the risk-free interest rate, $b(t,i) \equiv (b_1(t,i), \ldots, b_N(t,i))^\top \in \mathbb{R}^N$ be the vector of mean rates of return, and $\sigma(t,i) \equiv (\sigma_{nm}(t,i))_{N\times N} \in \mathbb{R}^{N\times N}$ be the volatility matrix. We assume the standard N-dimensional Brownian motion

$W(t) = (W_1(t), \ldots, W_N(t))^\top$ is independent of the Markov chain $\alpha(t)$. Furthermore, we assume that the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is generated by the Brownian motion $W(t)$ and the Markov chain $\alpha(t)$, augmented by all $\mathbb{P}$-null sets to satisfy the usual conditions.

The price process of the risk-free asset, denoted by $S_0(t)$, evolves according to the following ordinary differential equation (ODE),

$$dS_0(t) = r(t, \alpha(t))S_0(t)dt, \quad S_0(0) = s_0 > 0. \tag{2.1}$$

The price processes of the N risky assets, denoted by $S_n(t)$ for $n = 1, 2, \ldots, N$, are governed by the following stochastic differential equations (SDEs),

$$dS_n(t) = S_n(t)\left[b_n(t, \alpha(t))dt + \sum_{m=1}^{N} \sigma_{nm}(t, \alpha(t))dW_m(t)\right], \quad S_n(0) = s_n > 0. \tag{2.2}$$

For technical regularity, we assume that all the functions $r(t, i)$, $b_n(t, i)$, and $\sigma_{nm}(t, i)$ are measurable and uniformly bounded in t. Furthermore, we assume that the following nondegeneracy condition is satisfied for some $\delta > 0$,

$$\sigma(t, i)\sigma^\top(t, i) \geq \delta I, \quad \forall t \in [0, T], \quad \forall i \in \mathcal{M}, \tag{2.3}$$

where I is the $N \times N$ identity matrix.

Let $u(t) \equiv (u_1(t), u_2(t), \ldots, u_N(t))^\top \in \mathbb{R}^N$ denote the portfolio weight vector (the fraction of wealth invested in the N risky assets), where negative components indicate short selling and a sum exceeding 1 implies leveraging by borrowing at the risk-free rate. Let $c(t) \geq 0$ denote the consumption rate process. The wealth process $X(\cdot)$ under strategy (u, c) evolves according to the following SDE for $s \in [t, T]$,

$$\begin{cases} dX(s) = & \Big\{X(s)\big[r(s, \alpha(s)) + u^\top(s)(b(s, \alpha(s)) - r(s, \alpha(s))\mathbf{1})\big] - c(s)\Big\}ds \\ & +X(s)u^\top(s)\sigma(s, \alpha(s))dW(s), \\ X(t) = & x > 0, \quad \alpha(t) = i. \end{cases} \tag{2.4}$$

where, $\mathbf{1}$ is an N-dimensional vector of ones. For simplicity of notation, we introduce the market price of risk vector $\theta(t, i) := \sigma(t, i)^{-1}(b(t, i) - r(t, i)\mathbf{1})$. The wealth equation can be rewritten as

$$\begin{cases} dX(s) = & \Big\{X(s)\big[r(s, \alpha(s)) + u^\top(s)\sigma(s, \alpha(s))\theta(s, \alpha(s))\big] - c(s)\Big\}ds \\ & +X(s)u^\top(s)\sigma(s, \alpha(s))dW(s), \\ X(t) = & x > 0, \quad \alpha(t) = i. \end{cases} \tag{2.5}$$

The classical performance functional is defined as

$$J^{cl}(t, x, i; u, c) = \mathbb{E}_{t,x,i}\left[\int_t^T U_1(s, c(s))\, ds + U_2(X(T))\right], \tag{2.6}$$

where $\mathbb{E}_{t,x,i}[\cdot] := \mathbb{E}[\cdot \mid X(t) = x, \alpha(t) = i]$. We assume that the "consumption" and "investment" utility functions, $U_1 : [0, T] \times (0, \infty) \to \mathbb{R}$ and $U_2 : (0, \infty) \to \mathbb{R}$ respectively, satisfy the following standard conditions:

(i) For each $t \in [0, T]$, the mappings $x \to U_1(t, x)$ and $x \to U_2(x)$ are strictly increasing, strictly concave, and of class C^1 on $(0, \infty)$ (with x-derivatives denoted by $U_1'(t, x)$ and $U_2'(x)$ for each $(t, x) \in [0, T] \times (0, \infty)$) such that

$$\lim_{x\downarrow 0} U_1'(t, x) = +\infty, \quad \lim_{x\to\infty} U_1'(t, x) = 0,$$
$$\lim_{x\downarrow 0} U_2'(x) = +\infty, \quad \lim_{x\to\infty} U_2'(x) = 0;$$

(ii) The mappings $(t,x) \to U_1(t,x)$ and $(t,x) \to U_1^{'}(t,x)$ are continuous on $[0,T] \times (0,\infty)$.

Motivation for the exploratory framework. In the classical framework (2.5)-(2.6), the standard stochastic optimal control problem is to find an optimal admissible strategy (u^*, c^*) that maximizes the performance functional. While the classical regime-switching model provides a useful representation of financial reality compared to traditional models with deterministic coefficients, it generally assumes a known environment ([11]). In this classical setting, the investor typically requires prior knowledge of the Markov chain's transition rates, as well as the drift, volatility, and interest rate parameters for each market regime. In practice, however, these parameters can be challenging to estimate precisely and may remain partially unknown.

This practical consideration motivates the transition to an entropy-regularized reinforcement learning (ERRL) framework. By replacing deterministic controls with stochastic policies and incorporating a Shannon entropy regularizer, the agent is naturally encouraged to explore the state-action space ([16], [6]). This exploration allows the investor to dynamically learn effective consumption and investment strategies through interactions with the environment, thereby reducing the reliance on complete a priori knowledge of the underlying market dynamics.

2.2. Exploratory consumption-investment problem. In RL, exploration is achieved by substituting deterministic actions with probability distributions. In our model, the control variables are the consumption rate $c \in [0,\infty)$ and the risky asset weight vector $u \in \mathbb{R}^N$, yielding the joint action space $\mathcal{A} := [0,\infty) \times \mathbb{R}^N$.

To enable learning in an unknown environment, let $\boldsymbol{\pi} : (t,x,i) \in [0,T] \times (0,\infty) \times \mathcal{M} \to \boldsymbol{\pi}(\cdot \mid t,x,i) \in \mathcal{P}(\mathcal{A})$ be a given stochastic feedback control, where $\mathcal{P}(\mathcal{A})$ is the set of probability density functions defined on the action space $\mathcal{A}$. The investor selects their consumption and investment strategy at time s by sampling from the probability distribution $\boldsymbol{\pi}(c,u \mid s, X(s), \alpha(s))$.

Following the theoretical framework for continuous-time exploratory diffusions proposed by [16] and further extended by [2], we reformulate the classical problem using stochastic strategies. The detailed derivation of this exploratory state dynamics is provided in Appendix A. Under this exploratory framework, the state dynamics are governed by the corresponding regime-switching exploratory SDE

$$
\begin{cases}
dX(s) = & \Big(\displaystyle\int_{\mathcal{A}} \Big\{ X(s)\big[r(s,\alpha(s)) + u^\top \sigma(s,\alpha(s))\theta(s,\alpha(s))\big] - c \Big\} \\
& \qquad \times \boldsymbol{\pi}(c,u \mid s, X(s), \alpha(s))\, dc\, du \Big) ds \\
& + \Big\{ \displaystyle\int_{\mathcal{A}} X(s)^2 u^\top \sigma(s,\alpha(s)) \sigma^\top(s,\alpha(s)) u \\
& \qquad \times \boldsymbol{\pi}(c,u \mid s, X(s), \alpha(s))\, dc\, du \Big\}^{1/2} d\tilde{W}(s), \\
X(t) = & \quad x > 0, \quad \alpha(t) = i.
\end{cases} \tag{2.7}
$$

where $\tilde{W}(s)$ is a standard one-dimensional Brownian motion.

To encourage exploration, we add an entropy regularizer to the objective. The

exploratory performance functional is defined as

(2.8)
$$J(t,x,i;\boldsymbol{\pi}) = \mathbb{E}_{t,x,i}\Big[\int_t^T \Big(\int_{\mathcal{A}} U_1(s,c)\,\boldsymbol{\pi}(c,u \mid s,X(s),\alpha(s))\,dc\,du \\ -\lambda\int_{\mathcal{A}} \boldsymbol{\pi}(c,u \mid s,X(s),\alpha(s))\ln\boldsymbol{\pi}(c,u \mid s,X(s),\alpha(s))\,dc\,du\Big)ds + U_2(X(T))\Big],$$

where $\lambda > 0$ is the *temperature parameter*.

Remark 2.1. In ERRL, entropy encourages exploration by increasing policy randomness. Since the investor aims to maximize their expected performance in the proposed framework, the positive entropy term $-\lambda\int_{\mathcal{A}}\boldsymbol{\pi}\ln\boldsymbol{\pi}\,dc\,du$ acts as an exploration reward. The constant $\lambda > 0$ serves as the *temperature parameter*, influencing how quickly the investor updates their policy using new information and balancing the trade-off between exploitation (utility maximization) and the entropy incentive.

DEFINITION 2.2. *(Admissible policy). The policy $\boldsymbol{\pi} = \boldsymbol{\pi}(\cdot \mid t,x,i)$ is called admissible, and we denote the set of admissible policies by $\boldsymbol{\Pi}$, if:*

(*i*) *For all $(t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}$, $\boldsymbol{\pi}(\cdot \mid t,x,i) \in \mathcal{P}(\mathcal{A})$. Moreover, for every Borel set $B \in \mathcal{B}(\mathcal{A})$, the mapping $(t,x,i) \longmapsto \int_B \boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du$ is jointly measurable with respect to $\mathcal{B}([0,T])\otimes\mathcal{B}((0,\infty))\otimes 2^{\mathcal{M}}$. Consequently, for any $\{\mathcal{F}_s\}$-progressively measurable state process $(X(s),\alpha(s))$, the integrated drift and diffusion coefficients are $\{\mathcal{F}_s\}$-progressively measurable.*

(*ii*) *For any constants $0 < \varepsilon < K < \infty$, all $t \in [0,T]$, and all $i \in \mathcal{M}$, there exists a constant $C_{\varepsilon,K} > 0$ such that for all $x,y \in [\varepsilon,K]$,*

$$\int_{\mathcal{A}} \left(1 + c + u^\top u\right)\left|\boldsymbol{\pi}(c,u \mid t,x,i) - \boldsymbol{\pi}(c,u \mid t,y,i)\right| dc\,du \le C_{\varepsilon,K}|x-y|.$$

(*iii*) *For all $(t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}$ and some $p \ge 2$, there exists a constant $C > 0$ such that*

$$\int_{\mathcal{A}} \left|\ln\boldsymbol{\pi}(c,u \mid t,x,i)\right| \boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \le C\left(1 + |\ln x| + |x|^p\right).$$

(*iv*) *There exist constants $C > 0$ such that for all $(t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}$,*

$$\int_{\mathcal{A}} c\,\boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \le Cx, \quad \int_{\mathcal{A}} u^\top u\,\boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \le C.$$

Furthermore, to ensure non-degeneracy of exploration, there exists a constant $\underline{\epsilon} > 0$ such that

$$\int_{\mathcal{A}} u^\top u\,\boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \ge \underline{\epsilon}, \quad \forall (t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}.$$

PROPOSITION 2.3 (Well-posedness of the exploratory SDE). *For any admissible policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, the exploratory SDE (2.7) admits a unique strong solution $\{X(s)\}_{s\in[t,T]}$ given the initial condition $X(t) = x > 0$ and $\alpha(t) = i$. Furthermore, the solution remains strictly positive almost surely, i.e., $\mathbb{P}(X(s) > 0, \forall s \in [t,T]) = 1$.*

The proof can be found in Appendix B.

DEFINITION 2.4. *(Value function). The exploratory value function is defined as*

$$V(t,x,i) = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} J(t,x,i;\boldsymbol{\pi}). \tag{2.9}$$

2.3. Optimal exploratory policy. In this subsection, we employ the Dynamic Programming Principle (DPP) to elucidate the optimal exploratory policy for the consumption-investment problem. First, we derive the HJB equation for the exploratory value function (see Lemma 2.5), and then present the optimal policy as a functional of the solution to the HJB equation (see Theorem 2.8). Finally, we state a verification theorem (see Theorem 2.10) to guarantee that the solution to the HJB equation coincides with the true exploratory value function and that the generated policy is indeed optimal.

LEMMA 2.5 (HJB equation). *Suppose that for all $i \in \mathcal{M}$, $v(\cdot,\cdot,i) \in C^{1,2}([0,T]\times(0,\infty))$. Then, the formal HJB equation associated with the exploratory consumption-investment problem is given by*

$$\begin{aligned} -v_t(t,x,i) = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \Bigg[\int_{\mathcal{A}} H(t,x,i,v_x,v_{xx};c,u)\boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \\ -\lambda \int_{\mathcal{A}} \boldsymbol{\pi}(c,u \mid t,x,i)\ln \boldsymbol{\pi}(c,u \mid t,x,i)\,dc\,du \Bigg] + \sum_{j=1}^{D} q_{ij} v(t,x,j), \\ v(T,x,i) = U_2(x), \quad \forall x \in (0,\infty), \quad \forall i \in \mathcal{M}, \end{aligned} \tag{2.10}$$

where the classical Hamiltonian H is defined as

$$\begin{aligned} H(t,x,i,v_x,v_{xx};c,u) :=& U_1(t,c) + \left[x\left(r(t,i) + u^\top \sigma(t,i)\theta(t,i)\right) - c\right] v_x \\ &+ \frac{1}{2}x^2 u^\top \sigma(t,i)\sigma^\top(t,i) u v_{xx} \\ =& \left(\frac{1}{2}x^2 u^\top \sigma(t,i)\sigma^\top(t,i) u v_{xx} + x u^\top \sigma(t,i)\theta(t,i) v_x\right) \\ &+ (U_1(t,c) - c v_x) + x r(t,i) v_x. \end{aligned} \tag{2.11}$$

Proof. The proof follows the standard DPP for regime-switching diffusions [9], extended to the exploratory framework with entropy regularization [16]. ∎

LEMMA 2.6 (Monotonicity and concavity). *For each $i \in \mathcal{M}$, suppose $v(\cdot,\cdot,i) \in C^{1,2}([0,T)\times(0,\infty))$ is a finite-valued solution to the HJB equation (2.10). Then $v_x(t,x,i) > 0$, $v_{xx}(t,x,i) < 0$, for all $(t,x,i) \in [0,T)\times(0,\infty)\times\mathcal{M}$.*

The proof can be found in Appendix C.

For completeness, we define the mode as follows [12].

DEFINITION 2.7 (Mode). *The mode of a continuous random variable having density f is the value of x for which $f(x)$ is a maximum.*

To solve the supremum problem in the HJB equation (2.10), we employ the calculus of variations to maximize the objective functional with respect to the probability density $\boldsymbol{\pi}$, subject to the normalization constraint $\int_{\mathcal{A}} \boldsymbol{\pi}\,dc\,du = 1$.

THEOREM 2.8 (Optimal exploratory policy). *Suppose that the HJB equation (2.10) admits a finite-valued solution $v(t,x,i)$. Then the optimal exploratory policy $\boldsymbol{\pi}^*(c,u \mid t,x,i)$ separates into two independent distributions, $\boldsymbol{\pi}_c^*(c \mid t,x,i)$ and $\boldsymbol{\pi}_u^*(u \mid t,x,i)$. Specifically, the optimal investment policy is the multivariate Gaussian distribution*

$$\boldsymbol{\pi}_u^*(u \mid t,x,i) = \mathcal{N}\left(u \,\middle|\, -\frac{\sigma(t,i)^{-\top}\theta(t,i)v_x}{x v_{xx}}, \frac{\lambda(\sigma(t,i)\sigma^\top(t,i))^{-1}}{-x^2 v_{xx}}\right), \tag{2.12}$$

and the optimal consumption policy is the Gibbs distribution

$$\boldsymbol{\pi}_c^*(c \mid t,x,i) = \frac{\exp\left(\frac{1}{\lambda}[U_1(t,c) - cv_x]\right)}{\int_0^\infty \exp\left(\frac{1}{\lambda}[U_1(t,c) - cv_x]\right) dc}. \tag{2.13}$$

Proof. By Lemma 2.6, we have $v_x > 0$ and $v_{xx} < 0$.

Starting from the HJB equation (2.10), the maximizer $\boldsymbol{\pi}^*(c,u \mid t,x,i)$ solves

$$\sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \left\{ \int_{\mathcal{A}} H(t,x,i,v_x,v_{xx};c,u)\boldsymbol{\pi}\, dc\, du - \lambda \int_{\mathcal{A}} \boldsymbol{\pi} \ln \boldsymbol{\pi}\, dc\, du \right\},$$

subject to $\int_{\mathcal{A}} \boldsymbol{\pi}\, dc\, du = 1$. By the method of Lagrange multipliers,

$$\boldsymbol{\pi}^*(c,u \mid t,x,i) = \frac{\exp\left(\frac{1}{\lambda}H(t,x,i,v_x,v_{xx};c,u)\right)}{\int_{\mathcal{A}} \exp\left(\frac{1}{\lambda}H(t,x,i,v_x,v_{xx};c,u)\right)\, dc\, du}.$$

Using (2.11), the Hamiltonian can be decomposed as

$$H = xr(t,i)v_x + \big(U_1(t,c) - cv_x\big) + \frac{1}{2}x^2 v_{xx} u^\top \sigma(t,i)\sigma^\top(t,i)u + xv_x u^\top \sigma(t,i)\theta(t,i).$$

Hence the joint density factorizes as

$$\boldsymbol{\pi}^*(c,u \mid t,x,i) = \boldsymbol{\pi}_c^*(c \mid t,x,i)\boldsymbol{\pi}_u^*(u \mid t,x,i).$$

For the consumption policy,

$$\boldsymbol{\pi}_c^*(c \mid t,x,i) = \frac{\exp\left(\frac{1}{\lambda}[U_1(t,c) - cv_x]\right)}{\int_0^\infty \exp\left(\frac{1}{\lambda}[U_1(t,c) - cv_x]\right)\, dc}.$$

Since $v_x > 0$, the denominator is finite.

For the investment policy,

$$\boldsymbol{\pi}_u^*(u \mid t,x,i) \propto \exp\left(\frac{x^2 v_{xx}}{2\lambda} u^\top \sigma(t,i)\sigma^\top(t,i)u + \frac{xv_x}{\lambda} u^\top \sigma(t,i)\theta(t,i)\right).$$

Since $v_{xx} < 0$ and $\sigma(t,i)\sigma^\top(t,i)$ is positive definite, the quadratic term is strictly negative definite. Completing the square gives

$$\boldsymbol{\pi}_u^*(u \mid t,x,i) = \mathcal{N}\left(u \,\middle|\, -\frac{\sigma(t,i)^{-\top}\theta(t,i)v_x}{xv_{xx}}, \frac{\lambda(\sigma(t,i)\sigma^\top(t,i))^{-1}}{-x^2 v_{xx}}\right). \qquad \square$$

Remark 2.9. The modes of the probability density functions $\boldsymbol{\pi}_u^*(u \mid t,x,i)$ and $\boldsymbol{\pi}_c^*(c \mid t,x,i)$, as defined in Definition 2.7, coincide with the pointwise maximizers of the classical Hamiltonian H evaluated at the derivatives of the exploratory value function v. Specifically, these modes are given by

$$u^*_{\text{mode}} = -\frac{\sigma(t,i)^{-\top}\theta(t,i)v_x}{xv_{xx}}, \qquad c^*_{\text{mode}} = \left(U_1'(t,\cdot)\right)^{-1}(v_x).$$

These expressions are formally identical to the optimal deterministic controls in the classical framework.

While prior studies (see, e.g., [17]) have observed that the mean of the Gaussian exploratory distribution formally recovers the classical optimal control, our analysis highlights the role of the mode in characterizing the pointwise Hamiltonian maximizer. This distinction is particularly relevant for asymmetric distributions, where the mean and mode generally differ. To illustrate this observation, Section 3 considers a Gamma distribution for exploratory consumption, whose mode retains the functional form of the classical optimal consumption rule, whereas its mean differs from the mode.

Substituting the optimal exploratory policy $\boldsymbol{\pi}^*(c,u \mid t,x,i) = \boldsymbol{\pi}_c^*(c \mid t,x,i)\boldsymbol{\pi}_u^*(u \mid t,x,i)$ from Theorem 2.8 into the HJB equation (2.10), we obtain

$$\begin{aligned}
&\sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\left[\int_{\mathcal{A}} H(t,x,i,v_x,v_{xx};c,u)\boldsymbol{\pi}\,dc\,du - \lambda\int_{\mathcal{A}}\boldsymbol{\pi}\ln\boldsymbol{\pi}\,dc\,du\right]\\
=&\lambda\ln\left(\int_{\mathcal{A}}\exp\left(\frac{H(t,x,i,v_x,v_{xx};c,u)}{\lambda}\right)dc\,du\right)\\
=&xr(t,i)v_x - \frac{\theta(t,i)^\top\theta(t,i)v_x^2}{2v_{xx}} - \frac{\lambda N}{2}\ln(-x^2v_{xx}) - \frac{\lambda}{2}\ln\det(\sigma(t,i)\sigma^\top(t,i))\\
&+\frac{\lambda N}{2}\ln(2\pi\lambda) + \lambda\ln\left(\int_0^\infty\exp\left(\frac{U_1(t,c)-cv_x}{\lambda}\right)dc\right).
\end{aligned}$$

Therefore, $v(t,x,i)$ satisfies the following partial differential equation (PDE),

$$\begin{aligned}
-v_t(t,x,i) =&xr(t,i)v_x - \frac{\theta(t,i)^\top\theta(t,i)v_x^2}{2v_{xx}} - \frac{\lambda N}{2}\ln(-x^2v_{xx})\\
&-\frac{\lambda}{2}\ln\det(\sigma(t,i)\sigma^\top(t,i)) + \frac{\lambda N}{2}\ln(2\pi\lambda)\\
&+\lambda\ln\left(\int_0^\infty\exp\left(\frac{U_1(t,c)-cv_x}{\lambda}\right)dc\right) + \sum_{j=1}^D q_{ij}v(t,x,j),\\
v(T,x,i) =&U_2(x),\quad \forall x\in(0,\infty),\quad \forall i\in\mathcal{M}.
\end{aligned}\tag{2.14}$$

THEOREM 2.10 (Verification theorem). *Suppose that, for each $i\in\mathcal{M}$, there exists a finite-valued function $v(t,x,i)\in C^{1,2}([0,T]\times(0,\infty))$ satisfying the following conditions:*

(i) v solves the HJB equation (2.10) with terminal condition $v(T,x,i)=U_2(x)$;

(ii) there exist constants $C>0$ and $p>0$ such that $|v(t,x,i)|\le C\,(1+x^p+x^{-p})$ for all $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$;

(iii) for any admissible policy $\boldsymbol{\pi}\in\boldsymbol{\Pi}$, the local martingale term arising from the generalized Itô formula applied to $v(s,X(s),\alpha(s))$ is a supermartingale. Furthermore, the optimal exploratory policy $\boldsymbol{\pi}^*$ defined in (2.12) and (2.13) is admissible, and the corresponding local martingale term is a true martingale.

Then

$$v(t,x,i) = V(t,x,i) = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} J(t,x,i;\boldsymbol{\pi}),$$

and $\boldsymbol{\pi}^*$ is an optimal exploratory policy.

Proof. By Lemma 2.6, $v_x>0$ and $v_{xx}<0$. The result then follows from the standard verification argument for regime-switching stochastic control; see [14, 9]. ∎

3. Analysis under logarithmic utility. In this section, we specialize the general optimal exploratory consumption policy in Theorem 2.8 to logarithmic utility.

We derive the resulting Gamma distribution and analyze the convergence of the exploratory value function as the temperature parameter tends to zero. We also briefly discuss the extension to power utility.

Beyond their theoretical interest, the derived probability density functions offer useful insights for practical algorithm design. In particular, they provide a theoretical reference for selecting parametric policy distributions in actor networks, such as those used in soft actor-critic algorithms [3]. Choosing a distributional family that reflects the underlying economic structure may facilitate more stable and economically informed exploration in complex continuous action spaces.

COROLLARY 3.1 (Logarithmic utility). *Suppose the investor has a logarithmic preference with a subjective discount rate $\rho > 0$, i.e., $U_1(t,c) = e^{-\rho t}\ln(c)$, and the HJB equation (2.10) admits a finite-valued solution $v(t,x,i)$. Then the optimal exploratory consumption policy is given by*

$$\boldsymbol{\pi}_c^*(c \mid t,x,i) = \text{Gamma}\left(c \,\middle|\, \frac{e^{-\rho t}}{\lambda} + 1,\ \frac{v_x(t,x,i)}{\lambda}\right). \tag{3.1}$$

We now establish the connection between the classical consumption-investment problem and the exploratory consumption-investment problem under logarithmic utility, by examining the asymptotic behavior of the model as the temperature parameter $\lambda \to 0^+$.

PROPOSITION 3.2 (Convergence of the value function under logarithmic utility). *Suppose the investor has a logarithmic preference for consumption, i.e., $U_1(t,c) = e^{-\rho t}\ln(c)$. Let $V(t,x,i)$ denote the exploratory value function under temperature $\lambda > 0$, and $V^{cl}(t,x,i)$ denote the classical unregularized value function. As the exploration temperature $\lambda \to 0^+$, $V(t,x,i)$ converges uniformly to $V^{cl}(t,x,i)$ on any compact subset of the state space. Specifically, the approximation error is bounded by $\mathcal{O}(\lambda|\ln\lambda|)$.*

The proof can be found in Appendix D.

Building on the convergence of the value function established in Proposition 3.2, we can now examine the connection between the derived exploratory policy $\boldsymbol{\pi}^*(c,u \mid t,x,i)$ and the classical deterministic strategies. To do so, we analyze the asymptotic behavior of the exploratory policy distributions as the exploration temperature parameter $\lambda \to 0^+$. It is worth noting that these consistency results rely on the uniform convergence of the exploratory value function $V(t,x,i)$ to its classical counterpart $V^{cl}(t,x,i)$ as $\lambda \to 0^+$. Across both policies, the modes of the respective probability density functions formally coincide with the classical optimal strategies, though their distributional properties exhibit distinct behaviors.

- Investment policy ($\boldsymbol{\pi}_u^*$): The optimal portfolio distribution follows a multivariate Gaussian, where the mode and the mean naturally coincide. As $\lambda \to 0^+$, the covariance matrix given by $-\frac{\lambda(\sigma(t,i)\sigma^\top(t,i))^{-1}}{x^2 V_{xx}}$ continuously shrinks toward the zero matrix. Consequently, the distribution collapses via a Dirac delta measure onto its mean vector, which precisely recovers the classical optimal portfolio proportion

$$u^* = -\frac{\sigma(t,i)^{-\top}\theta(t,i)V_x}{xV_{xx}}.$$

- Consumption policy ($\boldsymbol{\pi}_c^*$): Under logarithmic utility, the exploratory consumption policy follows a Gamma distribution with shape parameter $\alpha =$

$\frac{e^{-\rho t}}{\lambda}+1$ and rate parameter $\beta=\frac{V_x}{\lambda}$. For any $\lambda>0$, the skewness introduced by entropy regularization causes the mode and the mean to differ. Specifically, while the expected value is $\mathbb{E}[c]=\frac{\alpha}{\beta}=\frac{e^{-\rho t}+\lambda}{V_x}$, the mode of the distribution precisely recovers the classical optimal consumption rate

$$c^*=\frac{\alpha-1}{\beta}=\frac{e^{-\rho t}}{V_x}.$$

As $\lambda\to 0^+$, the variance

$$\mathrm{Var}[c]=\frac{\lambda e^{-\rho t}+\lambda^2}{V_x^2}$$

vanishes. Consequently, the mean converges to the mode, and the distribution collapses via a Dirac delta measure centered at this classical deterministic consumption rate.

Together, these asymptotic properties demonstrate how the entropy-regularized exploratory framework smoothly transitions back to the classical framework as the exploration incentive fades away.

Remark 3.3 (Power utility). The optimal exploratory consumption policy can also be characterized under power utility. Suppose that

$$U_1(t,c)=e^{-\rho t}\frac{c^{1-\gamma}}{1-\gamma},\qquad \gamma>0,\quad \gamma\neq 1,$$

and that the HJB equation (2.10) admits a finite-valued solution $v(t,x,i)$. By Theorem 2.8, the optimal consumption density is

$$\boldsymbol{\pi}_c^*(c\mid t,x,i)=\frac{\exp\left(\frac{e^{-\rho t}c^{1-\gamma}}{\lambda(1-\gamma)}-\frac{v_x(t,x,i)}{\lambda}c\right)}{\int_0^\infty \exp\left(\frac{e^{-\rho t}z^{1-\gamma}}{\lambda(1-\gamma)}-\frac{v_x(t,x,i)}{\lambda}z\right)dz},\qquad c>0. \tag{3.2}$$

Unlike the logarithmic utility case, this density does not generally belong to a standard distribution family. Nevertheless, its mode is given explicitly by

$$c^*_{\mathrm{mode}}(t,x,i)=\left(e^{\rho t}v_x(t,x,i)\right)^{-1/\gamma},$$

which has the same functional form as the classical optimal consumption rule, with the exploratory value function replacing its unregularized counterpart.

4. q-Learning RL algorithm. While the optimal exploratory policy $\boldsymbol{\pi}^*(c,u\mid t,x,i)$ in Theorem 2.8 characterizes the exact solution to the exploratory optimal control problem, it relies on the market primitives $r(t,i)$, $b(t,i)$, $\sigma(t,i)$, and the transition rate matrix Q. In financial applications, these regime-switching dynamics are often difficult to estimate precisely or remain partially unknown a priori. To develop a model-free RL framework, we first define the associated action-value function (q-function) for the controller and establish its martingale characterization following continuous-time RL methodologies (cf. [5], [6], [19]).

4.1. q-Function and policy improvement. Based on the structure of the exploratory HJB equation (2.10), we introduce the q-function, which evaluates the action-value of a deterministic action $(c,u)\in\mathcal{A}$ before the agent acts according to its stochastic policy.

DEFINITION 4.1. *(q-Function). For a given $\boldsymbol{\pi}(\cdot \mid t,x,i) \in \mathbf{\Pi}$, let $J(t,x,i;\boldsymbol{\pi})$ be the corresponding exploratory payoff function. The q-function is defined as*

$$q(t,x,i;c,u;\boldsymbol{\pi}) := J_t(t,x,i;\boldsymbol{\pi}) + H(t,x,i,J_x,J_{xx};c,u) + \sum_{j=1}^{D} q_{ij} J(t,x,j;\boldsymbol{\pi}), \tag{4.1}$$

where H is the classical Hamiltonian defined in (2.11).

It immediately follows from the exploratory HJB equation that the q-function satisfies the following consistency condition

$$\int_{\mathcal{A}} \left(q(t,x,i;c,u;\boldsymbol{\pi}) - \lambda \ln \boldsymbol{\pi}(c,u \mid t,x,i)\right) \boldsymbol{\pi}(c,u \mid t,x,i)\, dc\, du = 0, \quad \forall (t,x,i). \tag{4.2}$$

LEMMA 4.2. *(Policy improvement). Let $\boldsymbol{\pi}(c,u \mid t,x,i) \in \mathbf{\Pi}$ be an arbitrarily given admissible policy. Suppose that the corresponding exploratory performance function $J(\cdot,\cdot,i;\boldsymbol{\pi}) \in C^{1,2}([0,T)\times(0,\infty)) \cap C^0([0,T]\times(0,\infty))$ for each $i \in \mathcal{M}$, and satisfies $J_x > 0$ and $J_{xx} < 0$ (i.e., J is strictly increasing and concave with respect to x). For any $(t,x,i) \in [0,T)\times(0,\infty)\times\mathcal{M}$, if the updated policy $\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)$ defined as*

$$\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i) = \frac{\exp\left(\frac{1}{\lambda} H(t,x,i,J_x,J_{xx};c,u)\right)}{\int_{\mathcal{A}} \exp\left(\frac{1}{\lambda} H(t,x,i,J_x,J_{xx};c',u')\right) dc'\, du'} \tag{4.3}$$

is admissible. Then, the exploratory performance functions under the specific policies $\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)$ and $\boldsymbol{\pi}(c,u \mid t,x,i)$ satisfy

$$J(t,x,i;\tilde{\boldsymbol{\pi}}) \geq J(t,x,i;\boldsymbol{\pi}), \quad \forall (t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}. \tag{4.4}$$

Proof. Fix $(t,x,i) \in [0,T]\times(0,\infty)\times\mathcal{M}$. Let $X_s^{\tilde{\boldsymbol{\pi}}}$ denote the wealth process driven by the updated policy $\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)$. For notational simplicity, define the exploratory drift and diffusion coefficients under policy $\tilde{\boldsymbol{\pi}}$ as

$$\begin{aligned} b^{\tilde{\boldsymbol{\pi}}}(t,x,i) =& xr(t,i) + x\int_{\mathcal{A}} u^\top \sigma(t,i)\theta(t,i)\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)\, dc\, du \\ & - \int_{\mathcal{A}} c\tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)\, dc\, du, \\ \sigma^{\tilde{\boldsymbol{\pi}}}(t,x,i) =& \left\{ x^2 \int_{\mathcal{A}} u^\top \sigma(t,i)\sigma^\top(t,i) u\, \tilde{\boldsymbol{\pi}}(c,u \mid t,x,i)\, dc\, du \right\}^{1/2}. \end{aligned}$$

Applying Itô's formula with regime switching to $J(s, X_s^{\tilde{\boldsymbol{\pi}}}, \alpha(s);\boldsymbol{\pi})$ along the path of $X_s^{\tilde{\boldsymbol{\pi}}}$, we have

$$\begin{aligned} & J(s, X_s^{\tilde{\boldsymbol{\pi}}}, \alpha(s);\boldsymbol{\pi}) - J(t,x,i;\boldsymbol{\pi}) \\ =& \int_t^s \Big[J_t(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-);\boldsymbol{\pi}) + b^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)) J_x(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-);\boldsymbol{\pi}) \\ & + \frac{1}{2}(\sigma^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)))^2 J_{xx}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-);\boldsymbol{\pi}) + \sum_{j=1}^{D} q_{\alpha(h-),j} J(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, j;\boldsymbol{\pi}) \Big] dh \\ & + \int_t^s J_x(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-);\boldsymbol{\pi}) \sigma^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)) d\tilde{W}(h) + \int_t^s dM_h^\alpha, \end{aligned}$$

where $\tilde{W}$ is the effective 1-dimensional Brownian motion, and M_h^α is the zero-mean local martingale associated with jumps of the continuous-time Markov chain α.

Define the localizing stopping times, for $n \geq 1$,

$$\tau_n := \inf\left\{ s \geq t : \int_t^s \left| J_x(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-); \boldsymbol{\pi}) \cdot \sigma^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)) \right|^2 dh \geq n \right\}.$$

Taking the conditional expectation of both sides, the martingale terms vanish, and we obtain

(4.5)

$$\begin{aligned}
&\mathbb{E}_{t,x,i}\left[J(s \wedge \tau_n, X_{s\wedge\tau_n}^{\tilde{\boldsymbol{\pi}}}, \alpha(s \wedge \tau_n); \boldsymbol{\pi}) \right] - J(t, x, i; \boldsymbol{\pi}) \\
=&\mathbb{E}_{t,x,i}\left[\int_t^{s\wedge\tau_n} \Big(J_t(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-); \boldsymbol{\pi}) + b^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)) J_x(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-); \boldsymbol{\pi}) \right. \right. \\
&\left. \left. + \frac{1}{2}(\sigma^{\tilde{\boldsymbol{\pi}}}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)))^2 J_{xx}(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-); \boldsymbol{\pi}) + \sum_{j=1}^D q_{\alpha(h-),j} J(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, j; \boldsymbol{\pi}) \Big) dh \right].
\end{aligned}$$

On the other hand, by the Feynman-Kac equation for the performance function under fixed policy $\boldsymbol{\pi}$, we have the equality for all $(t, x, i) \in [0, T) \times (0, \infty) \times \mathcal{M}$,

(4.6)

$$\begin{aligned}
-J_t(t, x, i; \boldsymbol{\pi}) = &\int_{\mathcal{A}} H(t, x, i, J_x, J_{xx}; c, u)\, \boldsymbol{\pi}(c, u \mid t, x, i)\, dc\, du \\
&- \lambda \int_{\mathcal{A}} \boldsymbol{\pi}(c, u \mid t, x, i) \ln \boldsymbol{\pi}(c, u \mid t, x, i)\, dc\, du + \sum_{j=1}^D q_{ij} J(t, x, j; \boldsymbol{\pi}).
\end{aligned}$$

By definition, $\tilde{\boldsymbol{\pi}}$ is the maximizer of the functional $\int_{\mathcal{A}} H\boldsymbol{\pi}\, dc\, du - \lambda \int_{\mathcal{A}} \boldsymbol{\pi} \ln \boldsymbol{\pi}\, dc\, du$ over all admissible policies. Thus for any policy $\boldsymbol{\pi}$,

$$\int_{\mathcal{A}} H\boldsymbol{\pi}\, dc\, du - \lambda \int_{\mathcal{A}} \boldsymbol{\pi} \ln \boldsymbol{\pi}\, dc\, du \leq \int_{\mathcal{A}} H\tilde{\boldsymbol{\pi}}\, dc\, du - \lambda \int_{\mathcal{A}} \tilde{\boldsymbol{\pi}} \ln \tilde{\boldsymbol{\pi}}\, dc\, du.$$

Substituting this into (4.6) evaluated along $(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-))$ yields

$$\begin{aligned}
&- J_t(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-); \boldsymbol{\pi}) \\
\leq &\int_{\mathcal{A}} H(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-), J_x, J_{xx}; c, u)\, \tilde{\boldsymbol{\pi}}(c, u \mid h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-))\, dc\, du \\
&- \lambda \int_{\mathcal{A}} \tilde{\boldsymbol{\pi}}(c, u \mid h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-)) \ln \tilde{\boldsymbol{\pi}}(c, u \mid h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, \alpha(h-))\, dc\, du \\
&+ \sum_{j=1}^D q_{\alpha(h-),j} J(h, X_{h-}^{\tilde{\boldsymbol{\pi}}}, j; \boldsymbol{\pi}).
\end{aligned} \tag{4.7}$$

Recall the classical Hamiltonian definition (2.11):

$$\int_{\mathcal{A}} H\tilde{\boldsymbol{\pi}}\, dc\, du = \int_{\mathcal{A}} U_1(h, c)\tilde{\boldsymbol{\pi}}\, dc\, du + b^{\tilde{\boldsymbol{\pi}}} J_x + \frac{1}{2}(\sigma^{\tilde{\boldsymbol{\pi}}})^2 J_{xx}.$$

Substituting it into (4.7) and re-arranging terms lead to

$$J_t + b^{\tilde{\boldsymbol{\pi}}} J_x + \frac{1}{2}(\sigma^{\tilde{\boldsymbol{\pi}}})^2 J_{xx} + \sum_{j=1}^D q_{\alpha(h-),j} J \geq -\int_{\mathcal{A}} U_1 \tilde{\boldsymbol{\pi}}\, dc\, du + \lambda \int_{\mathcal{A}} \tilde{\boldsymbol{\pi}} \ln \tilde{\boldsymbol{\pi}}\, dc\, du.$$

Plugging the above inequality into (4.5), we get

$$\begin{aligned} J(t,x,i;\boldsymbol{\pi}) \leq & \mathbb{E}_{t,x,i}\bigg[\int_t^{s\wedge\tau_n}\bigg(\int_{\mathcal{A}} U_1(h,c)\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))dc\,du \\ & -\lambda\int_{\mathcal{A}}\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))\ln\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))\,dc\,du\bigg)dh \\ & + J(s\wedge\tau_n, X^{\tilde{\boldsymbol{\pi}}}_{s\wedge\tau_n},\alpha(s\wedge\tau_n);\boldsymbol{\pi})\bigg]. \end{aligned}$$

Set $s=T$ and let $n\to\infty$. By the dominated convergence theorem and the terminal condition $J(T,X^{\tilde{\boldsymbol{\pi}}}_T,\alpha(T);\boldsymbol{\pi}) = U_2(X^{\tilde{\boldsymbol{\pi}}}_T)$, we derive that for all $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$,

$$\begin{aligned} J(t,x,i;\boldsymbol{\pi}) \leq & \mathbb{E}_{t,x,i}\bigg[\int_t^T\bigg(\int_{\mathcal{A}} U_1(h,c)\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))dc\,du \\ & -\lambda\int_{\mathcal{A}}\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))\ln\tilde{\boldsymbol{\pi}}(c,u\mid h,X^{\tilde{\boldsymbol{\pi}}}_{h-},\alpha(h-))\,dc\,du\bigg)dh + U_2(X^{\tilde{\boldsymbol{\pi}}}_T)\bigg] \\ = & J(t,x,i;\tilde{\boldsymbol{\pi}}). \end{aligned}$$

□

THEOREM 4.3. *(Policy improvement via q-function). Let $\boldsymbol{\pi}(c,u\mid t,x,i)\in\boldsymbol{\Pi}$ be an arbitrarily given admissible policy. Suppose that the corresponding exploratory performance function $J(\cdot,\cdot,i;\boldsymbol{\pi})\in C^{1,2}([0,T)\times(0,\infty))\cap C^0([0,T]\times(0,\infty))$ for each $i\in\mathcal{M}$, and satisfies $J_x>0$ and $J_{xx}<0$ (i.e., J is strictly increasing and concave with respect to x). For any $(t,x,i)\in[0,T)\times(0,\infty)\times\mathcal{M}$, if the updated policy $\tilde{\boldsymbol{\pi}}(c,u\mid t,x,i)$ defined as*

$$\tilde{\boldsymbol{\pi}}(c,u\mid t,x,i) = \frac{\exp\left(\frac{1}{\lambda}q(t,x,i;c,u;\boldsymbol{\pi})\right)}{\int_{\mathcal{A}}\exp\left(\frac{1}{\lambda}q(t,x,i;c',u';\boldsymbol{\pi})\right)dc'\,du'} \tag{4.8}$$

is admissible, then the exploratory performance function under the new policy satisfies

$$J(t,x,i;\tilde{\boldsymbol{\pi}})\geq J(t,x,i;\boldsymbol{\pi}). \tag{4.9}$$

Proof. The proof follows directly from Definition 4.1 and Lemma 4.2. □

4.2. Martingale characterization of the q-function. Following the ideas in [6, 19], this subsection specializes the martingale characterization for the continuous-time exploratory state process within a regime-switching setting.

LEMMA 4.4 (Characterization of the exploratory q-function). *Let an admissible policy $\boldsymbol{\pi}\in\boldsymbol{\Pi}$ be given, and suppose its value function $J(\cdot,\cdot,i;\boldsymbol{\pi})\in C^{1,2}([0,T]\times(0,\infty))\cap C([0,T]\times(0,\infty))$ for all $i\in\mathcal{M}$. Let a continuous function $\hat{q}:[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}\to\mathbb{R}$ be given. Then, we have*

(i) $\hat{q}(t,x,i,c,u) = q(t,x,i,c,u;\boldsymbol{\pi})$ for all $(t,x,i,c,u)\in[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}$ if and only if for all $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$, the following process

$$\begin{aligned} & J(s,X^{\boldsymbol{\pi}}(s),\alpha(s);\boldsymbol{\pi}) \\ & +\int_t^s\left[U_1(v,c^{\boldsymbol{\pi}}(v)) - \hat{q}(v,X^{\boldsymbol{\pi}}(v),\alpha(v),c^{\boldsymbol{\pi}}(v),u^{\boldsymbol{\pi}}(v))\right]dv, \end{aligned} \tag{4.10}$$

is an $(\{\mathcal{F}_s\}_{s\geq t},\mathbb{P})$*-martingale, where* $\{X^{\boldsymbol{\pi}}(s), t\leq s\leq T\}$ *is the solution to the exploratory SDE under* $\boldsymbol{\pi}$ *with initial conditions* $X^{\boldsymbol{\pi}}(t)=x$, $\alpha(t)=i$, *and* $(c^{\boldsymbol{\pi}}(v),u^{\boldsymbol{\pi}}(v))$ *is the action process sampled from* $\boldsymbol{\pi}(\cdot\mid v,X^{\boldsymbol{\pi}}(v),\alpha(v))$.

(ii) If $\hat{q}(t,x,i,c,u)=q(t,x,i,c,u;\boldsymbol{\pi})$ *for all* $(t,x,i,c,u)\in[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}$, *then given any other admissible policy* $\boldsymbol{\pi}'\in\boldsymbol{\Pi}$, *for all* $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$, *the following process*

$$J(s,X^{\boldsymbol{\pi}'}(s),\alpha(s);\boldsymbol{\pi}) + \int_t^s \Big[U_1(v,c^{\boldsymbol{\pi}'}(v)) - \hat{q}(v,X^{\boldsymbol{\pi}'}(v),\alpha(v),c^{\boldsymbol{\pi}'}(v),u^{\boldsymbol{\pi}'}(v))\Big]\,dv, \tag{4.11}$$

is an $(\{\mathcal{F}_s\}_{s\geq t},\mathbb{P})$*-martingale. Here,* $\{X^{\boldsymbol{\pi}'}(s), t\leq s\leq T\}$ *is the solution to the exploratory SDE under* $\boldsymbol{\pi}'$ *with* $X^{\boldsymbol{\pi}'}(t)=x$, $\alpha(t)=i$, *and* $(c^{\boldsymbol{\pi}'}(v),u^{\boldsymbol{\pi}'}(v))$ *is the action process sampled from* $\boldsymbol{\pi}'(\cdot\mid v,X^{\boldsymbol{\pi}'}(v),\alpha(v))$.

(iii) If there exists an admissible policy $\boldsymbol{\pi}'\in\boldsymbol{\Pi}$ *such that for all* $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$, *the process defined in* (4.11) *is an* $(\{\mathcal{F}_s\}_{s\geq t},\mathbb{P})$*-martingale with initial condition* $X^{\boldsymbol{\pi}'}(t)=x,\alpha(t)=i$, *then* $\hat{q}(t,x,i,c,u)=q(t,x,i,c,u;\boldsymbol{\pi})$ *for all* $(t,x,i,c,u)\in[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}$.

Moreover, in any of the three cases above, the q-function satisfies

$$\int_{\mathcal{A}}[q(t,x,i,c,u;\boldsymbol{\pi})-\lambda\ln\boldsymbol{\pi}(c,u\mid t,x,i)]\,\boldsymbol{\pi}(c,u\mid t,x,i)dc\,du=0. \tag{4.12}$$

The proof follows the same argument as that of Proposition 3.4 in [19] and is therefore omitted.

As in Theorem 3.5 of [19], we can strengthen Lemma 4.4 to jointly characterize the q-function and the value function associated with a given policy $\boldsymbol{\pi}$.

LEMMA 4.5 (Martingale characterization). *Let an admissible policy* $\boldsymbol{\pi}\in\boldsymbol{\Pi}$, *a function* $\hat{J}(\cdot,\cdot,i)\in C^{1,2}([0,T]\times(0,\infty))\cap C([0,T]\times(0,\infty))$ *for all* $i\in\mathcal{M}$, *and a continuous function* $\hat{q}:[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}\to\mathbb{R}$ *be given such that, for all* $(t,x,i,c,u)\in[0,T]\times(0,\infty)\times\mathcal{M}\times\mathcal{A}$,

$$\begin{cases}\hat{J}(T,x,i)=U_2(x),\\ \int_{\mathcal{A}}[\hat{q}(t,x,i,c,u)-\lambda\ln\boldsymbol{\pi}(c,u\mid t,x,i)]\,\boldsymbol{\pi}(c,u\mid t,x,i)\,dc\,du=0.\end{cases} \tag{4.13}$$

Then,

(i) $\hat{J}(t,x,i)$ *and* $\hat{q}(t,x,i,c,u)$ *are respectively the value function and the q function associated with* $\boldsymbol{\pi}$ *if and only if for all* $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$, *the following process*

$$\hat{J}(s,X^{\boldsymbol{\pi}}(s),\alpha(s))+\int_t^s[U_1(v,c^{\boldsymbol{\pi}}(v))-\hat{q}(v,X^{\boldsymbol{\pi}}(v),\alpha(v),c^{\boldsymbol{\pi}}(v),u^{\boldsymbol{\pi}}(v))]\,dv, \tag{4.14}$$

is an $(\{\mathcal{F}_s\}_{s\geq t},\mathbb{P})$*-martingale, where* $\{X^{\boldsymbol{\pi}}(s), t\leq s\leq T\}$ *is the solution to the exploratory SDE under* $\boldsymbol{\pi}$ *with initial conditions* $X^{\boldsymbol{\pi}}(t)=x$, $\alpha(t)=i$, *and* $(c^{\boldsymbol{\pi}}(v),u^{\boldsymbol{\pi}}(v))$ *is the action process sampled from* $\boldsymbol{\pi}(\cdot\mid v,X^{\boldsymbol{\pi}}(v),\alpha(v))$.

(ii) If $\hat{J}(t,x,i)$ *and* $\hat{q}(t,x,i,c,u)$ *are respectively the value function and the q-function associated with* $\boldsymbol{\pi}$, *then for any admissible policy* $\boldsymbol{\pi}'\in\boldsymbol{\Pi}$ *and all* $(t,x,i)\in[0,T]\times(0,\infty)\times\mathcal{M}$, *the following process*

$$\hat{J}(s,X^{\boldsymbol{\pi}'}(s),\alpha(s))+\int_t^s\Big[U_1(v,c^{\boldsymbol{\pi}'}(v))-\hat{q}(v,X^{\boldsymbol{\pi}'}(v),\alpha(v),c^{\boldsymbol{\pi}'}(v),u^{\boldsymbol{\pi}'}(v))\Big]\,dv, \tag{4.15}$$

is an $(\{\mathcal{F}_s\}_{s\geq t}, \mathbb{P})$*-martingale, where* $\{X^{\boldsymbol{\pi}'}(s), t \leq s \leq T\}$ *is the solution to the exploratory SDE under* $\boldsymbol{\pi}'$ *with* $X^{\boldsymbol{\pi}'}(t) = x, \alpha(t) = i$*, and* $(c^{\boldsymbol{\pi}'}(v), u^{\boldsymbol{\pi}'}(v))$ *is the action process sampled from* $\boldsymbol{\pi}'(\cdot \mid v, X^{\boldsymbol{\pi}'}(v), \alpha(v))$.

(iii) If there exists $\boldsymbol{\pi}' \in \mathbf{\Pi}$ *such that for all* $(t, x, i) \in [0, T] \times (0, \infty) \times \mathcal{M}$*, the process* (4.15) *is an* $(\{\mathcal{F}_s\}_{s\geq t}, \mathbb{P})$*-martingale with initial condition* $X^{\boldsymbol{\pi}'}(t) = x, \alpha(t) = i$*, then* $\hat{J}(t, x, i)$ *and* $\hat{q}(t, x, i, c, u)$ *are respectively the value function and the q-function associated with* $\boldsymbol{\pi}$.

Moreover, in any of the three cases above, if it holds further that $\boldsymbol{\pi}$ *takes the form of the Gibbs measure*

$$\boldsymbol{\pi}(c, u \mid t, x, i) = \frac{\exp\left(\frac{1}{\lambda}\hat{q}(t, x, i, c, u)\right)}{\int_{\mathcal{A}} \exp\left(\frac{1}{\lambda}\hat{q}(t, x, i, c', u')\right) dc'\, du'}, \tag{4.16}$$

assuming the normalizing constant is finite for all (t, x, i)*, then* $\boldsymbol{\pi}$ *is the optimal policy and* $\hat{J}$ *is the corresponding optimal value function.*

Remark 4.6. Lemma 4.5-(i) facilitates *on-policy* learning, where learning the q-function of a given target policy $\boldsymbol{\pi}$ is based on the data

$$\{(s, X_s^{\boldsymbol{\pi}}, \alpha_s, c_s^{\boldsymbol{\pi}}, u_s^{\boldsymbol{\pi}}), t \leq s \leq T\}$$

generated by $\boldsymbol{\pi}$ itself. On the other hand, Lemma 4.5-(ii) and -(iii) provide the theoretical foundation for *off-policy* learning, where learning the q-function of $\boldsymbol{\pi}$ is based on data generated by a different policy $\boldsymbol{\pi}'$, known as the behavior policy.

The continuous-time difference (CTD) error differential $d\delta_t$ for the q-function is given by the increment of the martingale

$$d\delta_t = d\hat{J}(t, X_t, \alpha_t) + (U_1(t, c_t) - \hat{q}(t, X_t, \alpha_t; c_t, u_t))\, dt. \tag{4.17}$$

By the martingale property, the expectation of the stochastic integral of this differential with respect to any predictable test process must be zero. This continuous-time orthogonality condition serves as the foundation for driving the parameter updates in our actor-critic architecture.

4.3. Algorithm design. We adopt a parameterized actor-critic framework that avoids direct integration over continuous action spaces. A key challenge in our regime-switching setting is that the optimal Gibbs policy derived from the entropy-regularized control problem involves a normalizing constant that, in general, cannot be computed in closed form, especially when the action space is jointly continuous over consumption c and investment u. Following the approach of [5, 19], we therefore parameterize both the value function and the policy within tractable families of distributions (e.g., Gaussian) and update the network parameters by minimizing the Kullback–Leibler (KL) divergence to the target Gibbs measure, rather than by evaluating the normalizing constant directly.

Let ϕ and ω denote the parameters of the value function $V_\phi(t, x, i)$ and the stochastic policy $\boldsymbol{\pi}_\omega(c, u \mid t, x, i)$, respectively. By Itô's lemma applied to V_ϕ along the grid-sampled trajectory, the q-function satisfies

$$\begin{aligned} q(t_k, X_{t_k}, \alpha_{t_k}; c_{t_k}, u_{t_k})\Delta t =& V(t_{k+1}, X_{t_{k+1}}, \alpha_{t_{k+1}}) - V(t_k, X_{t_k}, \alpha_{t_k}) \\ &+ U_1(t_k, c_{t_k})\Delta t + M(t_k, t_{k+1}), \end{aligned} \tag{4.18}$$

where $M(t_k, t_{k+1})$ is a martingale difference. Dropping it and replacing V by V_ϕ, the combination $(q - \lambda \ln \boldsymbol{\pi}_\omega)\Delta t$ arising in the policy gradient is approximated by the

empirical CTD error

$$\begin{aligned}\hat{\delta}_{t_k} = & V_\phi(t_{k+1}, X_{t_{k+1}}, \alpha_{t_{k+1}}) - V_\phi(t_k, X_{t_k}, \alpha_{t_k}) \\ & + \Big(U_1(t_k, c_{t_k}) - \lambda \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k})\Big)\Delta t,\end{aligned} \tag{4.19}$$

where, at the terminal time $t_K = T$, the value network is naturally anchored by the terminal wealth utility, satisfying $V_\phi(T, X_T, \alpha_T) = U_2(X_T)$.

4.3.1. Critic update. The critic parameters ϕ are updated to minimize the squared CTD error using a semi-gradient approach. Choosing the test function as $\xi_{t_k} = \partial_\phi V_\phi(t_k, X_{t_k}, \alpha_{t_k})$, the update rule is

$$\phi \leftarrow \phi + \alpha_\phi \partial_\phi V_\phi(t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}, \tag{4.20}$$

where $\alpha_\phi > 0$ denotes the learning rate for the critic network.

4.3.2. Actor update. To update the actor parameters ω, we minimize the Kullback-Leibler (KL) divergence between the parameterized policy $\boldsymbol{\pi}_\omega$ and the theoretically optimal Gibbs measure. Using the score function $\zeta_{t_k} = \partial_\omega \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k})$, the model-free update rule driven by the empirical CTD error is

$$\omega \leftarrow \omega + \alpha_\omega \partial_\omega \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}, \tag{4.21}$$

where $\alpha_\omega > 0$ represents the learning rate for the actor network.

Below, we summarize the implementation logic in two frameworks: an online incremental algorithm (Algorithm 4.1) and an offline episodic algorithm (Algorithm 4.2). We execute the stochastic policies on a uniform time grid with step size Δt such that $\mathbb{S} = \{0 = t_0 < t_1 < \cdots < t_K = T\}$.

Remark 4.7. In practical implementations of the actor update, if the theoretical functional form of the optimal policy distribution is known *a priori*, it is advantageous to parameterize $\boldsymbol{\pi}_\omega$ accordingly. For instance, in our continuous-time consumption-investment optimization problem, the theoretical optimal investment strategy u typically follows a Gaussian distribution, while the optimal consumption c follows a distribution defined on the positive half-line, such as a Gamma distribution. By designing the actor network to directly output the specific statistical parameters of these exact distributions (e.g., the mean and variance for the Gaussian distribution, and the shape and rate parameters for the Gamma distribution), we inject structural theoretical priors into the model. This explicit parameterization significantly reduces the hypothesis space compared to using a generic, black-box density estimator. Consequently, this structure not only accelerates the convergence of the actor-critic algorithm but also enhances training stability.

4.4. Numerical examples. In this subsection, we present numerical experiments to demonstrate the effectiveness of the proposed actor-critic q-learning algorithm for the exploratory regime-switching consumption-investment problem. To clarify the experimental setup, we explicitly distinguish between the parameters used to construct the **Environment simulator** (which governs the underlying dynamics but remains unknown to the learning agent) and the parameters required as known inputs for the learning algorithm.

Environment simulator parameters. The simulator is built upon a financial market with $D = 2$ regimes, representing a "bull" market (Regime 0) and a "bear" market (Regime 1). The hidden market parameters for the two regimes are set as follows.

Algorithm 4.1 Online Incremental q-Learning Algorithm

Input: Initial state $x_0 > 0$, initial regime i_0, horizon T, time step Δt, number of episodes N, number of mesh grids K, learning rates $\alpha_\phi, \alpha_\omega$, schedule $l(j)$, and temperature parameter λ.
Initialize: Critic parameters ϕ and actor parameters ω.
for episode $j = 1, 2, \ldots, N$ **do**
  Initialize $k = 0$. Observe initial state $X_{t_0} \leftarrow x_0$ and regime $\alpha_{t_0} \leftarrow i_0$.
  while $k < K$ **do**
    Sample action $(c_{t_k}, u_{t_k}) \sim \boldsymbol{\pi}_\omega(\cdot \mid t_k, X_{t_k}, \alpha_{t_k})$.
    Apply action (c_{t_k}, u_{t_k}) to the **Environment Simulator**; observe new wealth $X_{t_{k+1}}$, regime $\alpha_{t_{k+1}}$, and utility $r(t_k) = U_1(t_k, c_{t_k})$.
    Define the next-state evaluation:
    $v_{\text{next}} = U_2(X_{t_{k+1}})$ if $k = K - 1$; otherwise, $V_\phi(t_{k+1}, X_{t_{k+1}}, \alpha_{t_{k+1}})$.
    Compute the empirical CTD error:
    $\hat{\delta}_{t_k} = v_{\text{next}} - V_\phi(t_k, X_{t_k}, \alpha_{t_k}) + (r(t_k) - \lambda \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k}))\, \Delta t$
    Update Critic:
    $\phi \leftarrow \phi + l(j)\alpha_\phi \partial_\phi V_\phi(t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}$
    Update Actor:
    $\omega \leftarrow \omega + l(j)\alpha_\omega \partial_\omega \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}$
    Update $k \leftarrow k + 1$.
  end while
end for

- Regime 0 (Bull market): The risk-free rate $r(0) = 0.03$, expected return $\mu(0) = 0.08$, and volatility $\sigma(0) = 0.15$.
- Regime 1 (Bear market): The risk-free rate $r(1) = 0.01$, expected return $\mu(1) = 0.04$, and volatility $\sigma(1) = 0.25$.

The generator matrix of the continuous-time Markov chain is given by

$$Q = \begin{pmatrix} -0.5 & 0.5 \\ 1.0 & -1.0 \end{pmatrix}, \tag{4.22}$$

which implies that the market tends to recover from the bear state to the bull state faster than it transitions from the bull state to the bear state.

Learning algorithm inputs and settings. The agent requires several known inputs to execute the learning process. The finite time horizon is set to $T = 1.0$, which is divided into $K = 50$ mesh grids, yielding a time step of $\Delta t = 0.02$. The temperature parameter is inputted as $\lambda = 0.1$. To evaluate the agent's performance, the objective incorporates both an instantaneous consumption utility function and a terminal utility function. The specific mathematical forms of these utility functions will be explicitly defined in the subsequent subsections for different experimental scenarios.

To approximate the value function, the critic network employs a structured design using a LogSumExp formulation to guarantee monotonicity and concavity with respect to the wealth variable x. The neural networks for the actor and critic parameters (ω and ϕ) are trained using the Adam optimizer with initial learning rates $\alpha_\phi = \alpha_\omega = 10^{-3}$ over $N = 820$ episodes.

4.4.1. Numerical analysis for logarithmic utility function. We consider the case where both the instantaneous consumption utility and the terminal utility are logarithmic. Specifically, the consumption utility is given by $U_1(t, c) = e^{-\rho t} \ln(c)$

Algorithm 4.2 Offline Episodic q-Learning Algorithm

Input: Initial state $x_0 > 0$, initial regime i_0, horizon T, time step Δt, number of episodes N, number of mesh grids K, initial learning rates α_ϕ and α_ω, schedule $l(j)$, and temperature parameter λ.
Initialize: Critic parameters ϕ and actor parameters ω.
for episode $j = 1, 2, \ldots, N$ **do**
  Initialize $k = 0$. Observe initial state $X_{t_0} \leftarrow x_0$ and regime $\alpha_{t_0} \leftarrow i_0$.
  while $k < K$ **do**
    Sample action $(c_{t_k}, u_{t_k}) \sim \boldsymbol{\pi}_\omega(\cdot \mid t_k, X_{t_k}, \alpha_{t_k})$.
    Apply action (c_{t_k}, u_{t_k}) to the **Environment Simulator**.
    Observe new wealth state $X_{t_{k+1}}$, new regime $\alpha_{t_{k+1}}$, and instantaneous utility $r(t_k) = U_1(t_k, c_{t_k})$.
    Store transition tuple $(t_k, X_{t_k}, \alpha_{t_k}, c_{t_k}, u_{t_k}, r(t_k), X_{t_{k+1}}, \alpha_{t_{k+1}})$ in the episode buffer.
    Update $k \leftarrow k + 1$.
  end while
  for step $k = 0, 1, \ldots, K-1$ **do**
    Define the next-state evaluation:
    $v_{\text{next}} = U_2(X_{t_{k+1}})$ if $k = K-1$; otherwise, $V_\phi(t_{k+1}, X_{t_{k+1}}, \alpha_{t_{k+1}})$.
    Compute the empirical CTD error:
    $\hat{\delta}_{t_k} = v_{\text{next}} - V_\phi(t_k, X_{t_k}, \alpha_{t_k}) + (r(t_k) - \lambda \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k}))\,\Delta t$
  end for
  Compute the cumulative parameter updates:
  $\Delta\phi = \sum_{k=0}^{K-1} \partial_\phi V_\phi(t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}$
  $\Delta\omega = \sum_{k=0}^{K-1} \partial_\omega \ln \boldsymbol{\pi}_\omega(c_{t_k}, u_{t_k} \mid t_k, X_{t_k}, \alpha_{t_k}) \hat{\delta}_{t_k}$
  Apply parameter updates:
  $\phi \leftarrow \phi + l(j)\alpha_\phi \Delta\phi$; $\omega \leftarrow \omega + l(j)\alpha_\omega \Delta\omega$
end for

with a subjective discount rate $\rho = 0.5$, and the terminal utility is defined as $U_2(x) = \ln(x)$. To sample exploratory actions under this setting, the actor network models the investment strategy u using a Gaussian distribution and the consumption strategy c using a Gamma distribution.

Convergence of Losses. As depicted in Fig. 1, the squared CTD error of the Critic network decreases rapidly and stabilizes around zero. Concurrently, the Actor policy loss converges smoothly under the continuous-time policy entropy regularization. This dual stability indicates that the Critic effectively approximates the underlying HJB equation while the Actor maintains a well-balanced trade-off between exploration and exploitation.

Value Function Properties. Fig. 2 illustrates the learned value functions $V(0, x, i)$ for Regime 0 (Bull market) and Regime 1 (Bear market). Two economic features emerge as follows.

- Monotonicity and concavity: For both market regimes, $V(0, x, i)$ is strictly increasing and strictly concave with respect to wealth x, reflecting the investor's risk aversion.
- Regime superiority: The value function in Regime 0 consistently dominates that in Regime 1 across the entire wealth spectrum ($V(0, x, 0) > V(0, x, 1)$), confirming that the policy captures higher expected returns in favorable eco-

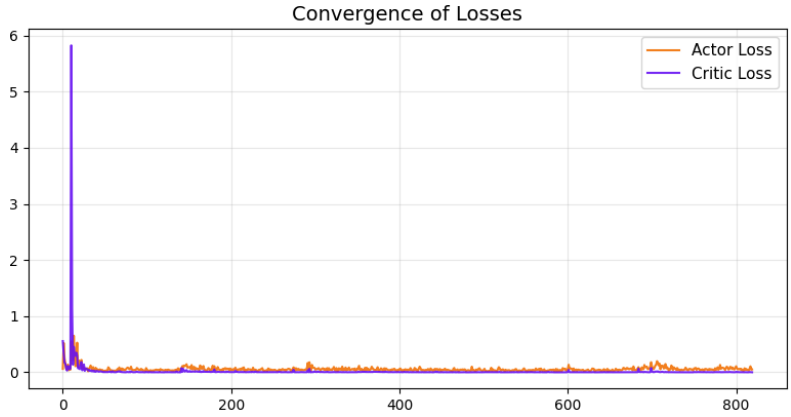


FIG. 1. *Convergence of Actor and Critic Losses. The figure illustrates a stable loss convergence under logarithmic consumption utility and logarithmic terminal utility.*

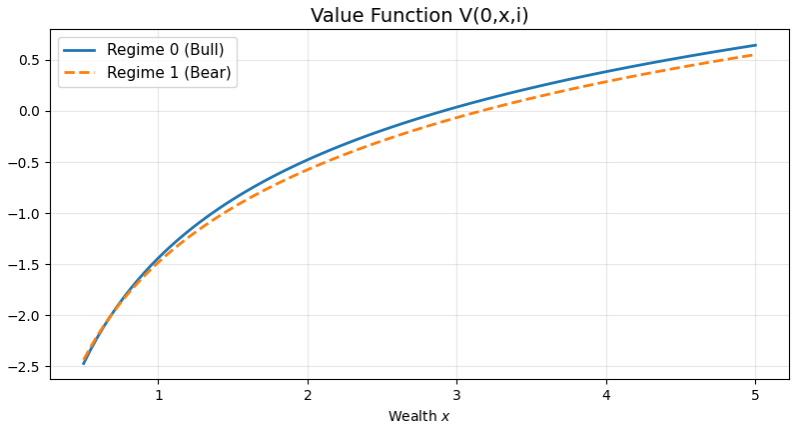


FIG. 2. *Value Function $V(0, x, i)$. The numerical performance demonstrates monotonic and concave value functions under logarithmic consumption utility and logarithmic terminal utility.*

nomic regimes.

Optimal Exploratory Strategy Distributions. Fig. 3 illustrates the optimal exploratory policies. To handle continuous action spaces, the Actor network parameterizes and learns the probability distributions of the actions. Notably, the classical deterministic optimal strategy corresponds to the mode (the peak) of these distributions rather than the mean.

- Investment strategy u (Gaussian): Modeled as $u \sim \mathcal{N}(\mu_u, \sigma_u^2)$. Because the Gaussian distribution is symmetric, its mode coincides with its mean μ_u. The peak for Regime 0 (Bull market, red line) is noticeably shifted to the right compared to Regime 1 (Bear market, black dashed line), indicating higher optimal capital allocation and risk exposure during economic prosperity.
- Consumption strategy c (Gamma): Modeled as $c \sim \text{Gamma}(\alpha_c, \beta_c)$ to enforce $c > 0$. Due to skewness, the mode distinctly differs from the expected value. Based on the distribution plot, the density peak for Regime 1 (Bear market, black dashed line) is higher and shifted to the left ($c \approx 0.85$) relative to Regime 0 (Bull market, red line, $c \approx 0.90$). This indicates a more

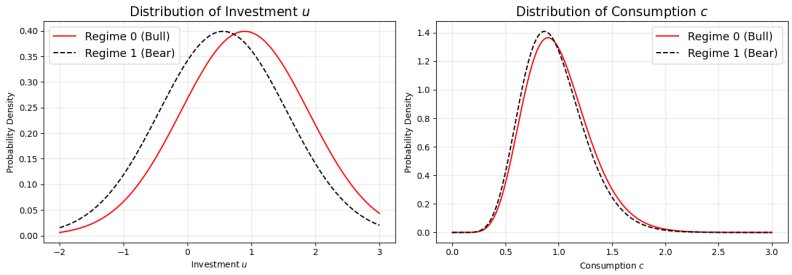


FIG. 3. *Exploratory Strategy Distributions of u and c. The policy parameterization uses a Gaussian distribution for the investment strategy u and a Gamma distribution for the consumption strategy c.*

concentrated and conservative strategy during the bear regime, restricting immediate consumption, whereas the high-return bull regime permits a slightly higher, more dispersed consumption profile.

4.4.2. Numerical examples for power utility function. To examine the applicability of the proposed algorithm beyond logarithmic utility, we consider power utility for both consumption and terminal wealth. Specifically, we set $U_1(t,c) = e^{-\rho t}c^{1-\gamma}/(1-\gamma)$ and $U_2(x) = x^{1-\gamma}/(1-\gamma)$, where $\rho = 0.5$ and the coefficient of relative risk aversion is $\gamma = 0.5$. This is the parameterization discussed in Remark 3.3.

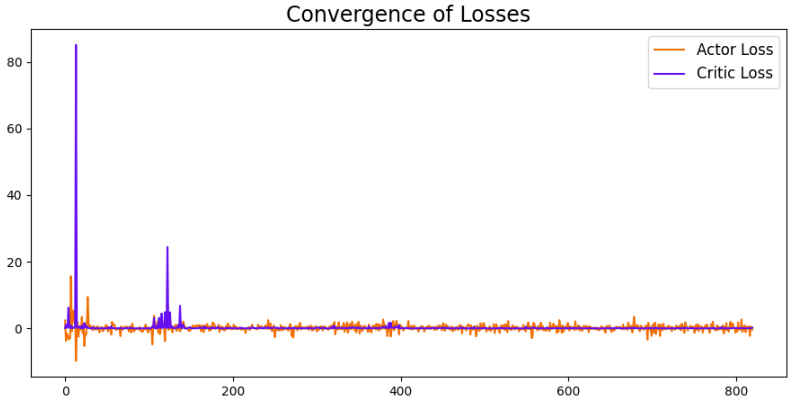


FIG. 4. *Convergence of Actor and Critic Losses. The figure illustrates stable convergence of the losses under power consumption utility and power terminal utility ($\gamma = 0.5$).*

Convergence of Losses. As illustrated in Fig. 4, both the Actor and Critic losses demonstrate overall convergence stability across the training episodes. Although a temporary disturbance in the Critic loss occurs around episode 130, the neural network rapidly self-corrects and stabilizes near zero. This rapid recovery highlights the numerical resilience and algorithmic stability of our continuous-time q-Learning framework in handling non-linear power utility specifications.

Value Function Properties. Fig. 5 depicts the learned value functions $V(0,x,i)$ at $t = 0$ under power utility ($\gamma = 0.5$). The learned values satisfy key economic properties.

- Monotonicity and concavity: Across both market regimes, $V(0,x,i)$ remains

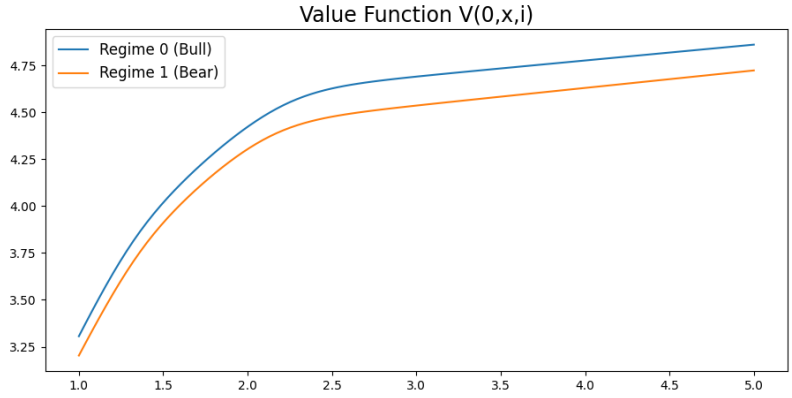


FIG. 5. *Value Function $V(0, x, i)$. The numerical performance demonstrates monotonic and concave value functions under power consumption utility and power terminal utility ($\gamma = 0.5$).*

strictly increasing and concave with respect to wealth x, consistent with the investor's persistent risk-averse behavior.

- Regime dominance: The value function in Regime 0 (Bull market, blue solid curve) dominates that in Regime 1 (Bear market, orange dashed curve) over the entire wealth grid ($V(0, x, 0) > V(0, x, 1)$), corroborating the higher expected reward potential in economic expansions.

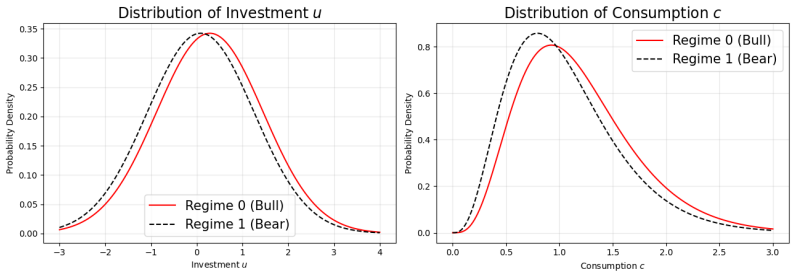


FIG. 6. *Exploratory Strategy Distributions of u and c. The policy parameterization uses a Gaussian distribution for the investment strategy u, while the empirical distribution of the consumption strategy c shares some structural similarities with a Gamma distribution under power utility ($\gamma = 0.5$).*

Optimal Exploratory Strategy Distributions. As shown in Fig. 6, the action probability distributions learned by the Actor network capture state-dependent investment and consumption behaviors.

- Investment strategy u (Gaussian): Parameterized as $u \sim \mathcal{N}(\mu_u, \sigma_u^2)$, the mode for Regime 0 (Bull market, red solid curve) is shifted rightward relative to Regime 1 (Bear market, black dashed curve). This demonstrates that the agent dynamically shifts its investment mode to allocate more capital toward the risky asset when market conditions are favorable.
- Consumption strategy c: The learned distribution of c satisfies positivity ($c > 0$) and exhibits an asymmetric profile that is somewhat reminiscent of a Gamma distribution. The density peak for Regime 0 shifts rightward (mode $\approx$ 1.0) compared to Regime 1 (mode ≈ 0.8). Unlike the logarithmic utility case, under power utility ($\gamma = 0.5$), the wealth effect induced by favorable regime

dynamics dominates, leading the investor to scale up immediate consumption in prosperous market states.

5. Conclusion. In this paper, we establish an entropy-regularized reinforcement learning (ERRL) framework for the continuous-time consumption-investment problem under a regime-switching diffusion model with general utility functions. By incorporating stochastic exploratory policies and Shannon entropy regularization into the stochastic optimal control formulation, we derive the explicit representations for the optimal exploratory portfolio and consumption strategies. Specifically, we show that the optimal investment strategy follows a multivariate Gaussian distribution, while the optimal consumption policy takes the form of a Gibbs measure, which reduces to a Gamma distribution under logarithmic utility. We also briefly illustrate the corresponding policy structure under power utility. Our analysis reveals that the modes of the optimal exploratory distributions retain the functional forms of the classical deterministic optimal controls, although their values need not coincide at a positive exploration temperature. Furthermore, under logarithmic utility, we establish the convergence of the exploratory value function to its classical unregularized counterpart as the exploration temperature parameter vanishes ($\lambda \to 0^+$).

To operate in model-free and unknown market environments, we develop a continuous -time q-learning actor-critic algorithm. We apply and specialize the existing theoretical results, proved the policy improvement theorem, and provide a martingale characterization for both on-policy and off-policy learning. Leveraging these theoretical results, we implement an actor-critic parameterization driven by continuous-time temporal difference (CTD) errors. Parameterizing the actor network based on the theoretical distribution families substantially accelerated the algorithm's convergence and enhanced numerical stability. Numerical experiments across regime-switching market environments validated our theoretical findings, demonstrating stable loss convergence, strict monotonicity and concavity of the learned value functions, and consistent regime superiority.

Several promising avenues remain open for future research. First, the proposed framework can be extended to incorporate market frictions such as transaction costs and short-selling constraints, or partially observable regime states. Second, generalizing the current framework to game-theoretic settings (such as stochastic differential games or mean-field games) presents a fascinating direction to study optimal strategies under competition or cooperation. Finally, exploring offline RL and model-based RL methodologies in continuous-time financial control is of great importance for improving sample efficiency and safe exploration in practical financial applications.

Acknowledgments. The authors would like to thank the handling editor and the anonymous reviewers for their insightful comments and constructive suggestions. The authors are also grateful to Xin Zhang of Southeast University for valuable advice.

References.

[1] L. Bo, Y. Huang, and X. Yu, *An extended Merton problem with relaxed benchmark tracking*, Mathematical Finance, 36 (2026), pp. 422-448.

[2] X. Gao, L. Li, and X.Y. Zhou, *Reinforcement learning for jump-diffusions, with financial applications*, Mathematical Finance, (2026), https://doi.org/10.1111/mafi.70027.

[3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*, in International Conference on Machine Learning, PMLR, 80 (2018), pp. 1861-1870.

[4] A.J. Heunis, *Utility maximization in a regime switching model with convex portfolio constraints and margin requirements: optimality relations and explicit solutions*, SIAM Journal on Control and Optimization, 53 (2015), pp. 2608-2656.
[5] Y. Jia and X.Y. Zhou, *Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach*, Journal of Machine Learning Research, 23 (2022), pp. 1-55.
[6] Y. Jia and X.Y. Zhou, *q-Learning in continuous time*, Journal of Machine Learning Research, 24 (2023), pp. 1-61.
[7] X. Li, Y. Wang, and X.Y. Zhou, *Dynamic mean-variance portfolio selection with no-shorting constraints and unknown investment opportunity sets*, (2026), https://arxiv.org/abs/2607.16625.
[8] G. Ma and S. Zhu, *Revisiting the Merton problem: from HARA to CARA utility*, Computational Economics, 59 (2022), pp. 651-688.
[9] X. Mao and C. Yuan, *Stochastic Differential Equations with Markovian Switching*, Imperial College Press, London, 2006.
[10] R. C. Merton, *Optimum consumption and portfolio rules in a continuous-time model*, in *Journal of Economic Theory*, 3 (1975), pp. 373-413.
[11] R. C. Merton, *On estimating the expected return on the market: An exploratory investigation*, Journal of Financial Economics, 8 (1980), pp. 323-361.
[12] S. M. Ross, *A First Course in Probability*, 8th ed., Pearson, London, 2014.
[13] S. E. Shreve and I. Karatzas, *Brownian Motion and Stochastic Calculus*, *Graduate Texts in Mathematics*, vol. 113, Springer, New York, 1991.
[14] L. R. Sotomayor and A. Cadenillas, *Explicit solutions of consumption-investment problems in financial markets with regime switching*, Mathematical Finance, 19 (2009), pp. 251-279.
[15] W. Tang, Y. P. Zhang, and X. Y. Zhou, *Exploratory HJB equations and their convergence*, SIAM Journal on Control and Optimization, 60 (2022), pp. 3191-3216.
[16] H. Wang, T. Zariphopoulou, and X.Y. Zhou, *Reinforcement learning in continuous time and space: A stochastic control approach*, Journal of Machine Learning Research, 21 (2020), pp. 1-34.
[17] H. Wang and X.Y. Zhou, *Continuous-time mean–variance portfolio selection: A reinforcement learning framework*, Mathematical Finance, 30 (2020), pp. 1273-1308.
[18] Y. Wang, J. Liu, K.F.C. Yiu, and T.K. Siu, *Consumption-portfolio optimization with regime-switching-modulated habit formation and jump diffusion*, Mathematics of Operations Research, (2026), https://doi.org/10.1287/moor.2024.0771.
[19] M. Zhang, X. Li, J. Xiong, and X. Zhang, *Continuous-time q-learning for Markov regime switching system under Tsallis entropy*, (2026), https://arxiv.org/pdf/2601.19299v2.
[20] M. Zhou, H. Zhou, and H. Meng, *Optimal consumption and investment in an incomplete market with hedgeable stochastic income*, European Journal of Operational Research, 330 (2026), pp. 311-325.

Appendix A. Exploratory dynamic system (2.7).

Consider the classical wealth dynamic system (2.5). We derive the corresponding exploratory wealth dynamics using the averaged infinitesimal generator associated with a randomized policy. For a fixed state $(s, x, i) \in [0, T] \times (0, \infty) \times \mathcal{M}$ and a deterministic action $(c, u) \in \mathcal{A}$, the drift and instantaneous variance of the classical

wealth process are given by

$$b(s,x,i;c,u) = x\big[r(s,i) + u^\top \sigma(s,i)\theta(s,i)\big] - c,$$
$$a(s,x,i;c,u) = x^2 u^\top \sigma(s,i)\sigma^\top(s,i)u.$$

For any sufficiently smooth test function $\varphi(s,x,i)$, the infinitesimal generator of the regime-switching wealth process under the fixed action (c,u) is

$$\mathcal{L}^{c,u}\varphi(s,x,i) = b(s,x,i;c,u)\varphi_x(s,x,i) + \frac{1}{2}a(s,x,i;c,u)\varphi_{xx}(s,x,i) + \sum_{j=1}^{D} q_{ij}\varphi(s,x,j).$$

Under the randomized policy $\boldsymbol{\pi}(c,u \mid s,x,i)$, the exploratory wealth dynamics are characterized by the averaged infinitesimal generator

$$\begin{aligned}\mathcal{L}^{\boldsymbol{\pi}}\varphi(s,x,i) &= \int_{\mathcal{A}} \mathcal{L}^{c,u}\varphi(s,x,i)\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du \\ &= b^{\boldsymbol{\pi}}(s,x,i)\varphi_x(s,x,i) + \frac{1}{2}a^{\boldsymbol{\pi}}(s,x,i)\varphi_{xx}(s,x,i) + \sum_{j=1}^{D} q_{ij}\varphi(s,x,j),\end{aligned}$$

where the averaged drift and instantaneous variance are respectively defined by

$$b^{\boldsymbol{\pi}}(s,x,i) = \int_{\mathcal{A}} \Big\{x\big[r(s,i) + u^\top\sigma(s,i)\theta(s,i)\big] - c\Big\}\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du,$$
$$a^{\boldsymbol{\pi}}(s,x,i) = \int_{\mathcal{A}} x^2 u^\top\sigma(s,i)\sigma^\top(s,i)u\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du.$$

The averaging of the generator corresponds to averaging the drift and quadratic variation characteristics of the classical wealth process under the randomized policy. Consequently, the exploratory wealth process can be represented by

$$\begin{cases} dX(s) = & \Big(\displaystyle\int_{\mathcal{A}} \Big\{X(s)\big[r(s,\alpha(s)) + u^\top\sigma(s,\alpha(s))\theta(s,\alpha(s))\big] - c\Big\} \\ & \qquad \times \boldsymbol{\pi}(c,u \mid s,X(s),\alpha(s))\,dc\,du\Big)ds \\ & + \Big\{\displaystyle\int_{\mathcal{A}} X(s)^2 u^\top\sigma(s,\alpha(s))\sigma^\top(s,\alpha(s))u \\ & \qquad \times \boldsymbol{\pi}(c,u \mid s,X(s),\alpha(s))\,dc\,du\Big\}^{1/2} d\tilde{W}(s), \\ X(t) = & \quad x > 0, \quad \alpha(t) = i. \end{cases}$$

Here, $\tilde{W}(s)$ is a standard one-dimensional Brownian motion. The regime-switching component remains governed by the generator matrix Q.

Appendix B. Proof of Proposition 2.3.

Proof. For notational simplicity, we denote the drift and diffusion coefficients of

the SDE (2.7) under policy $\boldsymbol{\pi}$ as $b_{\boldsymbol{\pi}}(s,x,i)$ and $\sigma_{\boldsymbol{\pi}}(s,x,i)$, respectively

$$\begin{aligned} b_{\boldsymbol{\pi}}(s,x,i) =& x\left(r(s,i)+\int_{\mathcal{A}} u^\top \sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du\right) \\ &-\int_{\mathcal{A}} c\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du, \\ \sigma_{\boldsymbol{\pi}}(s,x,i) =& \left\{x^2\int_{\mathcal{A}} u^\top\sigma(s,i)\sigma^\top(s,i)u\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du\right\}^{1/2}. \end{aligned}$$

Local Lipschitz continuity. Fix a compact interval $[\varepsilon,K]\subset(0,\infty)$ and let $x,y\in[\varepsilon,K]$. $C_{\varepsilon,K}>0$ denotes a constant that may change from line to line. By condition (ii) of Definition 2.2, together with the boundedness of σ and θ, we have

$$\begin{aligned} &\left|\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\left[\boldsymbol{\pi}(c,u\mid s,x,i)-\boldsymbol{\pi}(c,u\mid s,y,i)\right]dc\,du\right| \\ \le&\|\sigma\theta\|_\infty\int_{\mathcal{A}}|u|\;|\boldsymbol{\pi}(c,u\mid s,x,i)-\boldsymbol{\pi}(c,u\mid s,y,i)|\,dc\,du \\ \le&\|\sigma\theta\|_\infty\int_{\mathcal{A}}(1+u^\top u)\;|\boldsymbol{\pi}(x)-\boldsymbol{\pi}(y)|\,dc\,du\le C_{\varepsilon,K}|x-y|, \end{aligned}$$

where we used $|u|\le 1+u^\top u$, and $\|\sigma\theta\|_\infty:=\sup_{(t,j)\in[0,T]\times\mathcal{M}}\|\sigma(t,j)\theta(t,j)\|$ is the uniform supremum norm. Similarly,

$$\left|\int_{\mathcal{A}} c\left[\boldsymbol{\pi}(c,u\mid s,x,i)-\boldsymbol{\pi}(c,u\mid s,y,i)\right]dc\,du\right|\le C_{\varepsilon,K}|x-y|.$$

Moreover, by condition (iv) and the Cauchy–Schwarz inequality,

$$\left|\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du\right|\le\|\sigma\theta\|_\infty\left(\int_{\mathcal{A}} u^\top u\,\boldsymbol{\pi}\,dc\,du\right)^{1/2}\le C.$$

Therefore,

$$\begin{aligned} &|b_{\boldsymbol{\pi}}(s,x,i)-b_{\boldsymbol{\pi}}(s,y,i)| \\ =&\Big|x\Big(r(s,i)+\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du\Big) \\ &-y\Big(r(s,i)+\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u\mid s,y,i)\,dc\,du\Big) \\ &-\Big(\int_{\mathcal{A}} c\,\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du-\int_{\mathcal{A}} c\,\boldsymbol{\pi}(c,u\mid s,y,i)\,dc\,du\Big)\Big| \\ \le&|x-y|\left|r(s,i)+\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u\mid s,x,i)\,dc\,du\right| \\ &+y\left|\int_{\mathcal{A}} u^\top\sigma(s,i)\theta(s,i)\big[\boldsymbol{\pi}(c,u\mid s,x,i)-\boldsymbol{\pi}(c,u\mid s,y,i)\big]\,dc\,du\right| \\ &+\left|\int_{\mathcal{A}} c\left[\boldsymbol{\pi}(c,u\mid s,x,i)-\boldsymbol{\pi}(c,u\mid s,y,i)\right]dc\,du\right|\le C_{\varepsilon,K}|x-y|. \end{aligned}$$

Hence $b_{\boldsymbol{\pi}}(s,x,i)$ is locally Lipschitz continuous in x.

Next, for the diffusion coefficient, set

$$a_{\boldsymbol{\pi}}(s,x,i) := \int_{\mathcal{A}} u^{\top}\sigma(s,i)\sigma^{\top}(s,i)u\,\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du.$$

Since σ is uniformly bounded, condition (ii) gives

$$\begin{aligned}|a_{\boldsymbol{\pi}}(s,x,i) - a_{\boldsymbol{\pi}}(s,y,i)| &\le \|\sigma\sigma^{\top}\|_{\infty}\int_{\mathcal{A}} u^{\top}u\,|\boldsymbol{\pi}(c,u \mid s,x,i) - \boldsymbol{\pi}(c,u \mid s,y,i)|\,dc\,du\\ &\le C_{\varepsilon,K}|x-y|.\end{aligned}$$

Furthermore, by $\sigma(s,i)\sigma^{\top}(s,i) \ge \delta I$ and condition (iv),

$$a_{\boldsymbol{\pi}}(s,x,i) \ge \delta\int_{\mathcal{A}} u^{\top}u\,\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du \ge \delta\underline{\epsilon} > 0.$$

On the other hand, condition (iv) and the boundedness of σ imply that $a_{\boldsymbol{\pi}}(s,x,i) \le \overline{C}$ for some constant $\overline{C} > 0$. Therefore,

$$\begin{aligned}|\sigma_{\boldsymbol{\pi}}(s,x,i) - \sigma_{\boldsymbol{\pi}}(s,y,i)| &= \left|x\sqrt{a_{\boldsymbol{\pi}}(s,x,i)} - y\sqrt{a_{\boldsymbol{\pi}}(s,y,i)}\right|\\ &\le \sqrt{\overline{C}}\,|x-y| + y\left|\sqrt{a_{\boldsymbol{\pi}}(s,x,i)} - \sqrt{a_{\boldsymbol{\pi}}(s,y,i)}\right|.\end{aligned}$$

Using the identity

$$\left|\sqrt{a_{\boldsymbol{\pi}}(s,x,i)} - \sqrt{a_{\boldsymbol{\pi}}(s,y,i)}\right| = \frac{|a_{\boldsymbol{\pi}}(s,x,i) - a_{\boldsymbol{\pi}}(s,y,i)|}{\sqrt{a_{\boldsymbol{\pi}}(s,x,i)} + \sqrt{a_{\boldsymbol{\pi}}(s,y,i)}} \le \frac{C_{\varepsilon,K}}{2\sqrt{\delta\underline{\epsilon}}}|x-y|,$$

and $y \le K$, it follows that

$$|\sigma_{\boldsymbol{\pi}}(s,x,i) - \sigma_{\boldsymbol{\pi}}(s,y,i)| \le \left(\sqrt{\overline{C}} + \frac{KC_{\varepsilon,K}}{2\sqrt{\delta\underline{\epsilon}}}\right)|x-y|.$$

Thus, $\sigma_{\boldsymbol{\pi}}(s,x,i)$ is locally Lipschitz continuous in x on $(0,\infty)$.

Linear growth condition. Since r, σ, and θ are uniformly bounded, there exist constants $M_r, M_\sigma, M_\theta > 0$ such that $|r(s,i)| \le M_r$, $\|\sigma(s,i)\| \le M_\sigma$, $\|\theta(s,i)\| \le M_\theta$. By the definition of $b_{\boldsymbol{\pi}}$ and the triangle inequality,

$$\begin{aligned}|b_{\boldsymbol{\pi}}(s,x,i)| \le& x|r(s,i)| + x\left|\int_{\mathcal{A}} u^{\top}\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du\right|\\ &+ \int_{\mathcal{A}} c\,\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du.\end{aligned}$$

By the Cauchy–Schwarz inequality and condition (iv),

$$\begin{aligned}\left|\int_{\mathcal{A}} u^{\top}\sigma(s,i)\theta(s,i)\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du\right| &\le M_\sigma M_\theta\int_{\mathcal{A}} |u|\,\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du\\ &\le M_\sigma M_\theta\left(\int_{\mathcal{A}} u^{\top}u\,\boldsymbol{\pi}(c,u \mid s,x,i)\,dc\,du\right)^{1/2}\\ &\le M_\sigma M_\theta\sqrt{C}.\end{aligned}$$

Therefore, using again condition (iv),

$$|b_{\boldsymbol{\pi}}(s,x,i)| \le \Big(M_r + M_\sigma M_\theta \sqrt{C} + C\Big)\, x \le K_1(1+x)$$

for some constant $K_1 > 0$. Similarly,

$$\begin{aligned}|\sigma_{\boldsymbol{\pi}}(s,x,i)| &= x\left\{\int_{\mathcal{A}} u^\top \sigma(s,i)\sigma^\top(s,i)u\, \boldsymbol{\pi}(c,u \mid s,x,i)\, dc\, du\right\}^{1/2} \\ &\le x\left\{M_\sigma^2 \int_{\mathcal{A}} u^\top u\, \boldsymbol{\pi}(c,u \mid s,x,i)\, dc\, du\right\}^{1/2} \le M_\sigma\sqrt{C}\, x \le K_1(1+x).\end{aligned}$$

Hence, both $b_{\boldsymbol{\pi}}$ and $\sigma_{\boldsymbol{\pi}}$ satisfy the linear growth condition.

By the standard existence and uniqueness theorem for SDEs with locally Lipschitz coefficients and linear growth, the exploratory SDE (2.7) admits a unique maximal strong solution on $[t,\tau_0)$, where

$$\tau_0 := \inf\{s \ge t : X(s) = 0\}$$

is the first hitting time of zero.

Strict positivity. It remains to show that $\tau_0 > T$ almost surely. For $n \ge 1$ define

$$\tau_n := \inf\{s \ge t : X(s) \le 1/n\}.$$

For all sufficiently large n, $1/n < x$, so $X(s) > 0$ on $[t,\tau_n]$. On this interval, Itô's formula applied to $\ln X(s)$ gives

$$d\ln X(s) = \widetilde{\mu}(s)\, ds + \widetilde{\nu}(s)\, d\widetilde{W}(s),$$

where

$$\widetilde{\mu}(s) := \frac{b_{\boldsymbol{\pi}}(s,X(s),\alpha(s))}{X(s)} - \frac{1}{2}a_{\boldsymbol{\pi}}(s,X(s),\alpha(s)), \qquad \widetilde{\nu}(s) := \sqrt{a_{\boldsymbol{\pi}}(s,X(s),\alpha(s))}.$$

Since

$$\frac{b_{\boldsymbol{\pi}}(s,x,i)}{x} = r(s,i) + \int u^\top \sigma\theta\, \boldsymbol{\pi}(\cdot \mid s,x,i)\, dc\, du - \frac{1}{x}\int c\, \boldsymbol{\pi}(\cdot \mid s,x,i)\, dc\, du,$$

condition (iv) and Cauchy–Schwarz yield $|b_{\boldsymbol{\pi}}(s,x,i)/x| \le M_r + M_\sigma M_\theta\sqrt{C} + C =: M_b$. Together with $a_{\boldsymbol{\pi}} \le \overline{C}$, this gives

$$|\widetilde{\mu}(s)| \le M_b + \tfrac{1}{2}\overline{C} =: M', \qquad |\widetilde{\nu}(s)| \le \sqrt{\overline{C}} \qquad \text{a.s. on } [t,\tau_n].$$

Integrating on $[t, s\wedge\tau_n]$,

$$\ln X(s\wedge\tau_n) = \ln x + \int_t^{s\wedge\tau_n} \widetilde{\mu}(r)\, dr + M_n(s), \qquad M_n(s) := \int_t^{s\wedge\tau_n} \widetilde{\nu}(r)\, d\widetilde{W}(r).$$

Here M_n is a continuous square-integrable martingale with

$$\langle M_n\rangle_s = \int_t^{s\wedge\tau_n} a_{\boldsymbol{\pi}}(r,X(r),\alpha(r))\, dr \le \overline{C}(T-t),$$

so Doob's L^2 inequality gives the n-uniform bound

$$\mathbb{E}\Big[\sup_{s\in[t,T]}|M_n(s)|^2\Big] \le 4\,\mathbb{E}[\langle M_n\rangle_T] \le 4\overline{C}(T-t).$$

On the event $\{\tau_n \le T\}$, continuity gives $X(\tau_n) = 1/n$, hence

$$-\ln n = \ln x + \int_t^{\tau_n} \widetilde{\mu}(r)\,dr + M_n(\tau_n).$$

Since $|\int_t^{\tau_n} \widetilde{\mu}\,dr| \le M'(T-t)$, we get $M_n(\tau_n) \le -\ln n - \ln x + M'(T-t)$. Thus, for $L_n := \ln n + \ln x - M'(T-t) > 0$,

$$\{\tau_n \le T\} \subseteq \Big\{\sup_{s\in[t,T]}|M_n(s)| \ge L_n\Big\}.$$

By Chebyshev's inequality,

$$\mathbb{P}(\tau_n \le T) \le \frac{\mathbb{E}[\sup_{s\in[t,T]}|M_n(s)|^2]}{L_n^2} \le \frac{4\overline{C}(T-t)}{\big[\ln n + \ln x - M'(T-t)\big]^2} \longrightarrow 0.$$

Since $\tau_n \uparrow \tau_0$ (by continuity of X), the event $\{\tau_0 \le T\}$ equals $\bigcap_{n\ge n_0}\{\tau_n \le T\}$ up to finitely many initial indices, so continuity of probability yields

$$\mathbb{P}(\tau_0 \le T) = \lim_{n\to\infty}\mathbb{P}(\tau_n \le T) = 0.$$

Hence $\mathbb{P}(X(s) > 0,\ \forall s \in [t,T]) = 1$.

Finally, the linear growth condition excludes finite-time explosion, while the preceding argument excludes hitting the boundary 0. Hence the maximal strong solution extends uniquely to the whole interval $[t,T]$. □

Appendix C. Proof of Lemma 2.6.

Proof. Fix (t,x,i). By the entropy variational formula, finiteness of the right-hand side of (2.10) requires

$$\int_{\mathcal{A}} \exp\left(\frac{1}{\lambda}H(t,x,i,v_x,v_{xx};c,u)\right)\,dc\,du < \infty.$$

Using (2.11), the action-dependent part of the Hamiltonian separates as

$$U_1(t,c) - cv_x + \frac{1}{2}x^2 v_{xx}u^\top\sigma\sigma^\top u + xv_x u^\top\sigma\theta.$$

First, if $v_x < 0$, then $-cv_x \to \infty$ linearly as $c \to \infty$. If $v_x = 0$, since $U_1(t,\cdot)$ is increasing,

$$\int_0^\infty \exp\left(\frac{U_1(t,c)}{\lambda}\right)dc = \infty.$$

Hence $v_x > 0$.

Next, by the nondegeneracy condition $\sigma\sigma^\top \ge \delta I$, if $v_{xx} > 0$, the quadratic term in u is positive and dominates the linear term as $\|u\| \to \infty$, so the integral over $\mathbb{R}^N$ is infinite. If $v_{xx} = 0$, the integrand is either exponential in a nonzero linear function of u, or constant in u when $v_x\sigma\theta = 0$; in either case its integral over $\mathbb{R}^N$ is infinite. Therefore $v_{xx} < 0$. Thus v is strictly increasing and strictly concave in x. □

Appendix D. Proof of Proposition 3.2.

Proof. Define

$$A_0(t) = 1 + \int_t^T e^{-\rho s}\, ds, \quad A_\lambda(t) = A_0(t) + \lambda(T-t).$$

Under logarithmic utility, the classical and exploratory HJB equations admit solutions of the form

$$V^{cl}(t,x,i) = A_0(t)\ln x + B_0(t,i), \quad V(t,x,i) = A_\lambda(t)\ln x + B_\lambda(t,i).$$

Substituting these expressions into the corresponding HJB equations yields

$$-\partial_t B_k(t,i) = F_k(t,i) + \sum_{j=1}^{D} q_{ij} B_k(t,j), \qquad B_k(T,i) = 0, \tag{D.1}$$

for $k \in \{0,\lambda\}$, where

$$F_0(t,i) = A_0(t)\left(r(t,i) + \frac{1}{2}\theta(t,i)^\top\theta(t,i)\right) + e^{-\rho t}\left(\ln\frac{e^{-\rho t}}{A_0(t)} - 1\right),$$

and

$$\begin{aligned} F_\lambda(t,i) = {} & A_\lambda(t)\left(r(t,i) + \frac{1}{2}\theta(t,i)^\top\theta(t,i)\right) + \frac{\lambda N}{2}\ln\frac{2\pi\lambda}{A_\lambda(t)} \\ & - \frac{\lambda}{2}\ln\det\big(\sigma(t,i)\sigma^\top(t,i)\big) + \lambda\ln\Gamma\left(1 + \frac{e^{-\rho t}}{\lambda}\right) + (e^{-\rho t} + \lambda)\ln\frac{\lambda}{A_\lambda(t)}. \end{aligned}$$

Since $A_\lambda(t) \geq 1$, the boundedness and nondegeneracy assumptions on the market coefficients ensure that the associated optimal Gaussian and Gamma policies are admissible. Standard verification arguments, together with the required martingale integrability conditions, identify these solutions with the respective value functions.

Next, by Stirling's formula, uniformly in $t \in [0,T]$,

$$\begin{aligned} & \lambda\ln\Gamma\left(1 + \frac{e^{-\rho t}}{\lambda}\right) + (e^{-\rho t} + \lambda)\ln\frac{\lambda}{A_\lambda(t)} \\ & \quad = e^{-\rho t}\left(\ln\frac{e^{-\rho t}}{A_\lambda(t)} - 1\right) + \frac{\lambda}{2}\ln\lambda + \mathcal{O}(\lambda). \end{aligned}$$

Since $A_\lambda(t) - A_0(t) = \mathcal{O}(\lambda)$ uniformly in t, it follows that

$$F_\lambda(t,i) - F_0(t,i) = \frac{N+1}{2}\lambda\ln\lambda + \mathcal{O}(\lambda), \tag{D.2}$$

uniformly in $(t,i) \in [0,T] \times \mathcal{M}$.

Writing $B_k(t) = (B_k(t,1), \ldots, B_k(t,D))^\top$, the solution of (D.1) satisfies

$$B_k(t) = \int_t^T e^{Q(s-t)} F_k(s)\, ds.$$

Since Q is a Markov chain generator, $e^{Q(s-t)}$ is a stochastic matrix. Therefore, (D.2) implies

$$B_\lambda(t,i) - B_0(t,i) = \frac{N+1}{2}(T-t)\lambda\ln\lambda + \mathcal{O}(\lambda),$$

uniformly in (t, i). Consequently,

$$V(t,x,i) - V^{cl}(t,x,i) = \lambda(T-t)\ln x + \frac{N+1}{2}(T-t)\lambda\ln\lambda + \mathcal{O}(\lambda).$$

For any compact interval $[\varepsilon, R] \subset (0,\infty)$, the function $\ln x$ is uniformly bounded. Hence, for sufficiently small $\lambda > 0$,

$$\sup_{(t,x,i)\in[0,T]\times[\varepsilon,R]\times\mathcal{M}} |V(t,x,i) - V^{cl}(t,x,i)| \le C_{\varepsilon,R}\lambda|\ln\lambda|.$$

Since $\lambda|\ln\lambda| \to 0$ as $\lambda \to 0^+$, the desired locally uniform convergence follows. □