Merton's Problem with Recursive Perturbed Utility

Min Dai
Department of Applied Mathematics and School of Accounting and Finance, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. mindai@polyu.edu.hk

Yuchao Dong
Key Laboratory of Intelligent Computing and Applications (Tongji University), Ministry of Education and School of Mathematical Sciences, Tongji University, Shanghai 200092, China. ycdong@tongji.edu.cn

Yanwei Jia
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. yanweijia@cuhk.edu.hk

Xun Yu Zhou
Department of Industrial Engineering and Operations Research & Data Science Institute, Columbia University, New York, NY 10027, USA. xz2574@columbia.edu

The classical Merton investment problem predicts deterministic, state-dependent portfolio *rules*; however, laboratory and field evidence suggests that individuals often prefer randomized decisions leading to stochastic and noisy choices. Fudenberg et al. (2015) develop the additive perturbed utility theory to explain the preference for randomization in the static setting, which, however, becomes ill-posed or intractable in the dynamic setting. We introduce the *recursive* perturbed utility (RPU), a special stochastic differential utility that incorporates an entropy-based preference for randomization into a recursive aggregator. RPU endogenizes the intertemporal trade-off between utilities from randomization and bequest via a discounting term dependent on past accumulated randomization, thereby avoiding excessive randomization and yielding a well-posed problem. In a general Markovian incomplete market with CRRA preferences, we prove that the RPU-optimal portfolio policy (in terms of the risk exposure ratio) is Gaussian and can be expressed in closed form, independent of wealth. Its variance is inversely proportional to risk aversion and stock volatility, while its mean is based on the solution to a partial differential equation. Moreover, the mean is the sum of a myopic term and an intertemporal hedging term (against market incompleteness) that intertwines with policy randomization. Finally, we carry out an asymptotic expansion in terms of the perturbed utility weight to show that the optimal mean policy deviates from the classical Merton policy at first order, while the associated relative wealth loss is of a higher order, quantifying the financial cost of the preference for randomization.

Key words: Merton's problem; preference for randomization; recursive perturbed utility; biased stochastic policy

1. Introduction

Merton's continuous-time investment problem (Merton, 1969) is a cornerstone of intertemporal portfolio choice. It provides a tractable framework for characterizing optimal investment decisions and has been extended to a variety of preference specifications with rich economic implications. A common conclusion arising from these models is that a rational, risk-averse investor follows a *deterministic* function of time and state, also called a portfolio policy or rule.[1] Thus, one can predict investor behavior deterministically based on these models. However, experimental and empirical evidence suggests that people sometimes favor or even crave "randomized" and noisy choices, with examples ranging from sushi omakase[2] and "blind boxes"[3] to portfolio allocations. In other words, the same individual does not necessarily repeat the same choice when faced with the same problem. This leads to *stochastic* choice and randomized policies. Fudenberg et al. (2015) propose and develop the *additive perturbed utility* (APU) for *static* choice problems, which is expected utility plus a perturbation over choice probabilities. This perturbation captures an inherent desire to randomize, consistent also with the behavioral literature (Hey and Carbone, 1995; Dwenger et al., 2018; Miao and Zhong, 2018; Permana, 2020; Agranov and Ortoleva, 2017, 2025). The most widely used perturbation is the *entropy* function, which produces the logit/Luce stochastic choice when the choice set is finite (Luce et al., 1959; Anderson et al., 1992).[4]

In this paper, we extend APU to the *dynamic* setting inherent for Merton's problem. Now that the control is continuum-valued, this extension from static to dynamic is by no means straightforward. Indeed, randomizing portfolios does not generate direct monetary payoffs; rather, the perturbation term is a non-monetary component of the preference in addition to the bequest utility (of the terminal wealth). Because portfolio choices are made continuously over time, the agent accrues a *flow* of utility from randomization, analogous to the flow utility from consumption in the classical Merton consumption-investment problem.[5] With this in mind, our first result shows that the dynamic counterpart of APU in the Merton setting is *ill-posed* when risk aversion is not sufficiently strong: the agent can drive the payoff to infinity by increasing the choice variance without bound. Two forces, specific to the dynamic setting, are behind this pathology. First, with a continuous control, the entropy-based randomization can deliver unbounded perturbation payoffs, unlike the bounded entropy in finite choice sets. Second, the intertemporal trade-off between randomization and investment fundamentally differs from the familiar trade-off between consumption and investment. Consumption sacrifices future wealth growth; randomization, by contrast, leaves the *expected* growth rate unchanged while increasing volatility. Risk aversion, if too low, is therefore insufficient to deter excessive randomization. Moreover, even when risk aversion is sufficiently large to restore well-posedness, the additive model fails to yield tractable solutions—even in the Black-Scholes setting with constant-relative-risk-aversion (CRRA) preferences—making it difficult to characterize and analyze optimal policies.

These failures highlight more nuanced intertemporal interactions between utilities from randomization and from investment than the APU can capture. We therefore propose the *recursive perturbed utility* (RPU), inspired by recursive utility developed by Epstein and Zin (1989) in discrete time and by Duffie and Epstein (1992) in continuous time. While recursive utility was introduced to separate risk aversion from the elasticity of intertemporal substitution in consumption, our goal is to *endogenize* the effect of the entropy perturbation. The form of RPU resembles the Uzawa utility (Uzawa, 1968), which captures habit formation through an endogenous discount rate affected by past consumption.[6] In our setting, past randomization influences the agent's current preference in the form of time preference: the perturbation induces *endogenous discounting* that dampens the flow of utility from further randomization as the latter accumulates over time. In other words, RPU gives rise to a history-dependent time preference on randomization, such that the more or longer the investor has randomized in the past, the less weight she places on current randomization.

We study Merton's problem in an incomplete Markovian stock market with exogenous stochastic factors (e.g., Wachter 2002; Liu 2007; Chacko and Viceira 2005) and CRRA utilities within the

RPU framework featuring entropy perturbations. We prove that the optimal portfolio (in terms of the risk exposure ratio) policy must follow a Gaussian distribution that is dependent on the stochastic factors but independent of wealth. The variance of this policy admits a closed-form expression, which is inversely proportional to both relative risk aversion and the instantaneous variance of stock returns. Thus, higher risk aversion or higher market volatility reduces choice noise. Meanwhile, the optimal mean is dependent only on time and factor, characterized by a partial differential equation (PDE). In general, the mean decomposes into a myopic term (that depends on risk aversion and the stock's instantaneous return-risk trade-off but is *independent* of the randomization) and an intertemporal hedging term (that captures the hedging need due to market incompleteness and its correlation with the randomization). In complete markets or when the stock and factor move independently, the hedging term vanishes and the mean optimal policy is purely myopic, coinciding with the classical Merton solution. A notable special case is log utility (unit relative risk aversion), in which intertemporal hedging is absent, and RPU collapses to APU yielding a mean optimal policy identical to the Merton benchmark.[7]

In general, however, RPU biases the optimal mean relative to the classical Merton benchmark (without randomization). We quantify this effect via an asymptotic expansion in the perturbation weight (called the "*temperature*"). The deviation of the optimal mean from the classical Merton policy is first order in temperature, while the associated relative wealth loss from adopting the RPU-optimal policy (versus the classical benchmark) is of higher order. This in turn provides a transparent measure of the financial cost associated with the preference for randomized choices.

It is worth noting that the perturbed utility with entropy function has been widely used in the reinforcement learning (RL) literature, albeit with a different name "*entropy regularization*", as seen in works such as Ziebart et al. (2008); Haarnoja et al. (2018); Zhao et al. (2019) for the discrete-time setting and Wang et al. (2020) for the continuous-time one, among many others. The resulting optimal policies take the form of a Gibbs measure. However, the purposes of introducing randomized policies and entropy regularization for RL are conceptually and fundamentally different: the aim is to design algorithms to solve decision-making problems in a data-driven way, often without knowledge of the environment's probabilistic structure. The entropy term is added to the reward function as an explicit incentive to encourage *exploration* by randomizing choices, with the ultimate goal to balance exploration (learning) and exploitation (optimization).

By contrast, our paper follows Fudenberg et al. (2015) and employs entropy to capture humans' intrinsic preference for randomization.[8] We study and characterize optimal behaviors under this perturbed utility paradigm by assuming the investor is rational and possesses full knowledge of the market model. Here, stochastic choices do not arise from the need to explore, but rather from a genuine preference for randomization.[9]

Different motivations aside, the mathematical framework for incorporating stochastic policies and entropy functions in continuous time is premised upon the same notion of *relaxed controls* introduced by Wang et al. (2020). Numerous subsequent works have laid the mathematical foundations for randomized controls in continuous time. For example, Bender and Thuan (2026) and Jia et al. (2026) confirm that state processes obtained by randomizing decisions at discrete time grids converge weakly to the so-called exploratory state process in Wang et al. (2020) as the sampling frequency approaches infinity. Tang et al. (2022) establish the well-posedness of the Hamilton-Jacobi-Bellman equation associated with the exploratory formulation in Wang et al. (2020) by a viscosity solution approach. The framework has also been extended to mean-field games (Guo et al., 2022), optimal stopping problems (Dianetti et al., 2025; Dai et al., 2025b), risk-sensitive problems (Jia, 2024), and time-inconsistent problems (Dai et al., 2023). All these papers adopt the additive form, directly adding the entropy function as a running reward. Other forms of regularization or perturbation functions in continuous time have been studied by Han et al. (2023), but they remain restricted to the additive form.

The remainder of this paper is organized as follows. Section 2 presents the market setup, formulates Merton's problem and motivates the recursive additive entropy utility. Section 3 proves the optimality of Gaussian randomized policy and discusses conditions under which the policy is biased or unbiased. The section further provides an asymptotic analysis of the impact of the primary temperature parameter. Finally, Section 4 concludes. Additional results and discussions are provided in the appendix.

2. Problem Formulation

2.1. Market Environment and Investment Objective

Consider a financial market with two available investment assets: a risk-free bond offering a constant interest rate r and a risky stock (or market index). The stock price process S_t evolves according to the stochastic differential equation (SDE):

$$\frac{\mathrm{d}S_t}{S_t} = \mu(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t,\ S_0 = s_0, \tag{1}$$

where B denotes a one-dimensional Brownian motion. The instantaneous return rate $\mu_t \equiv \mu(t, X_t)$ and volatility $\sigma_t \equiv \sigma(t, X_t)$ both depend on an observable stochastic market factor process X_t. The dynamics of X_t are given by

$$\mathrm{d}X_t = m(t, X_t)\mathrm{d}t + \nu(t, X_t)[\rho \mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}\tilde{B}_t],\ X_0 = x_0, \tag{2}$$

where $\tilde{B}$ is an independent one-dimensional Brownian motion, and $\rho \in (-1, 1)$ is a constant representing the correlation between the stock return and the market factor changes. As a result, the market is generally incomplete. We focus on the Markovian setting, where the functions $\mu(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$, $m(\cdot,\cdot)$, and $\nu(\cdot,\cdot)$ are deterministic and continuous in both t and x such that the SDEs (1)-(2) admit a unique weak solution. This market environment is sufficiently general to encompass many widely studied incomplete market models as special cases—for example, the Gaussian mean return model and the stochastic volatility model examined in Wachter (2002), Liu (2007), Chacko and Viceira (2005); Dai et al. (2021), among others.[10]

The investor's actions are represented as a scalar-valued non-anticipative process $a = \{a_t\}_{t\in[0,T)}$, where a_t denotes the fraction of total wealth allocated to the stock at time t. The associated self-financing wealth process W^a evolves according to the SDE:

$$\frac{\mathrm{d}W_t^a}{W_t^a} = [r + (\mu(t, X_t) - r)a_t]\mathrm{d}t + \sigma(t, X_t)a_t\mathrm{d}B_t, \quad W_0^a = w_0. \tag{3}$$

It is important to note that the solvency constraint $W_t^a \geqslant 0$ almost surely for all $t \in [0,T]$, is automatically satisfied for any square integrable a. The classical Merton investment problem seeks to maximize the expected bequest utility of terminal wealth W_T^a by selecting an appropriate strategy a:

$$\max_a \mathbb{E}\left[U(W_T^a)\right], \tag{4}$$

subject to the dynamics (2)-(3), where $U(\cdot)$ is the utility function.

In the main text, we focus on the constant relative risk aversion (CRRA) utility, given by

$$U(w) = \frac{w^{1-\gamma} - 1}{1-\gamma},$$

where $1 \neq \gamma > 0$ is the relative risk aversion parameter. The additive constant "$-\frac{1}{1-\gamma}$" does not affect the preference reflected; we include it here for two reasons. First, the form converges to the log utility as $\gamma \to 1$. Second, including the constant will properly reflect the magnitude of the bequest utility when we combine it with the preference for randomization in Section 2.3. The case for the constant absolute risk aversion (CARA) utility function will be discussed in Appendix B.

2.2. Preference for Randomization

As it stands, the classical Merton problem leads to a *deterministic* optimal (feedback) policy. The agent has no appetite for randomization because it would increase the uncertainty in wealth, which is unfavored by risk aversion. Therefore, to capture the preference for randomization, we need to add reward for randomization explicitly into the objective function. Moreover, with randomized choices, the resulting law of motion of the wealth process described by (3) will also be fundamentally altered because a_t will now be randomly sampled from a distribution that is independent of the market randomness $(B_t, \tilde{B}_t)$. The resulting mathematical formulation is put forward and coined as the "exploratory formulation" for general stochastic control problems by Wang et al. (2020) in the reinforcement learning (RL) context. We now adapt that formulation to the Merton problem.

The central idea of this formulation is to describe the portfolio rules using a probability distribution for randomization. Specifically, suppose the investor selects her action (portfolio) at time

t by sampling from a probability distribution $\boldsymbol{\pi}_t$, where $\{\boldsymbol{\pi}_t\}_{t\in[0,T]} =: \boldsymbol{\pi}$ is a distribution-valued process referred to as a randomized *control*. Under such a control, the exploratory or randomized dynamics of the wealth process are given by

$$\frac{\mathrm{d}W_t^{\boldsymbol{\pi}}}{W_t^{\boldsymbol{\pi}}} = [r + (\mu(t,X_t) - r)\,\mathrm{Mean}(\boldsymbol{\pi}_t)]\,\mathrm{d}t + \sigma(t,X_t)\left[\mathrm{Mean}(\boldsymbol{\pi}_t)\mathrm{d}B_t + \sqrt{\mathrm{Var}(\boldsymbol{\pi}_t)}\mathrm{d}\bar{B}_t\right],\ W_0^{\boldsymbol{\pi}} = w_0, \quad (5)$$

where $\bar{B}$ is another Brownian motion, independent of both B and $\tilde{B}$, representing the additional randomness introduced into the wealth process due to randomization. Equation (5) shows that randomization effectively raises the volatility of the resulting wealth process, and $\mathrm{Mean}(\boldsymbol{\pi}_t)$ plays a similar role as a_t in the classical dynamics (3). Intuitively, $W^{\boldsymbol{\pi}}$ can be viewed as the "average" of infinitely many wealth processes generated by portfolio processes repeatedly sampled from the *same* randomized control $\boldsymbol{\pi}$. Another interpretation of $W^{\boldsymbol{\pi}}$ is the weak limit of wealth processes under piecewise constant portfolios where the portfolios are sampled from $\boldsymbol{\pi}$ only at discrete time points, as the mesh size of the sampling grid tends to zero; see Bender and Thuan (2026) and Jia et al. (2026). The derivation of (5) is analogous to that in Dai et al. (2023) and Dai et al. (2025a); see a detailed explanation in Appendix A of Dai et al. (2025a) and Jia et al. (2026) for how (5) can be viewed as the limit of sampling randomized choice at a high frequency.

Second, to describe the preference for randomization, we adopt the entropy function to measure the level of randomness associated with a distribution $\boldsymbol{\pi}$, denoted as $\mathcal{H}(\boldsymbol{\pi}) = -\int_{\mathbb{R}} \boldsymbol{\pi}(a)\log\boldsymbol{\pi}(a)\mathrm{d}a$ (known as the differential entropy).[11] For Merton's problem, the simplest way seems to be just adding running APU from randomization to the bequest utility, leading to the maximization of the following objective:

$$\mathbb{E}\left[\int_0^T \lambda\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + U(W_T^{\boldsymbol{\pi}})\right], \quad (6)$$

where $\lambda > 0$ is the *temperature parameter* representing the importance of randomization relative to the bequest utility, and $\{\lambda\mathcal{H}(\boldsymbol{\pi}_t)\}_{t\in[0,T]}$ is the flow of utility from randomization. The form (6) is a direct adaptation of the APU arising from the static setting of Fudenberg et al. (2015); it has also been chosen in the RL context by many algorithms as an incentive for exploration (e.g., Ziebart et al. 2008; Haarnoja et al. 2018 among many others).

As it turns out, however, it is difficult to analyze the randomized Merton problem under the objective (6), where a closed-form solution is unavailable even for the Black-Scholes market with the CRRA utility (see Appendix A.1 for explanations). Even worse, Proposition 1 below shows that the problem with the objective (6) is ill-posed if risk aversion is not sufficiently strong.

PROPOSITION 1. *If* $\gamma \in (0,1)$*, then the problem with the objective functional* (6)*, subject to the wealth dynamics* (5)*, is ill-posed with an infinite optimal value.*

Proof. When $\gamma \in (0,1)$, $U(w) > -\frac{1}{1-\gamma}$ for any $w > 0$. We consider a simple policy $\boldsymbol{\pi}_t = \mathcal{N}(0,v)$, where $v > 0$ is a constant. Then the objective function in (6) is greater than $\frac{\lambda T}{2}\log(2\pi e v) - \frac{1}{1-\gamma}$. Letting $v \to \infty$ causes the objective value to diverge to infinity. Q.E.D.

The intuition of Proposition 1 is that when the investor is not sufficiently risk-averse, there is not enough deterrence from over-randomization while the utility from the entropy term can be unbounded. Indeed, the proof of the proposition indicates that the problem becomes ill-posed whenever the bequest utility U is bounded from below. Hence, a naïve APU is not appropriate in the Merton setting to yield reasonable economic predictions. To remedy this problem, we introduce a different type of perturbed utility for randomization.

2.3. Recursive Perturbation Utility

Let a function $\lambda(\cdot,\cdot) > 0$ be exogenously given, called a *primary* temperature function (it will be taken as a constant in Section 3.3). Define by $J^{\boldsymbol{\pi}} = \{J_t^{\boldsymbol{\pi}}\}_{t\in[0,T]}$ the following *recursive* (entropy) perturbed utility (RPU) under a given randomized control $\boldsymbol{\pi}$, which is an $\{\mathcal{F}_t\}_{t\in[0,T]}$-adapted process satisfying

$$J_t^{\boldsymbol{\pi}} = \mathbb{E}\left[\int_t^T \lambda(s,X_s)[(1-\gamma)J_s^{\boldsymbol{\pi}}+1]\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + U(W_T^{\boldsymbol{\pi}})\middle|\mathcal{F}_t\right], \tag{7}$$

where $W^{\boldsymbol{\pi}}$ is the wealth process under $\boldsymbol{\pi}$, determined by (2) and (5), and $\{\mathcal{F}\}_{t\in[0,T]}$ is the filtration generated by $\mathbf{B} := (B,\bar{B},\tilde{B})^\top$.

The recursive form (7) effectively weighs the entropy utility endogenously: The term $\lambda(t,X_t)\,[(1-\gamma)J_t^{\boldsymbol{\pi}}+1]$ can be viewed as a *utility-dependent* weight on randomization determined in a recursive way. Notice that when $\gamma = 1$ (corresponding to the log utility function), the weight on the entropy term becomes independent of the utility, and (7) further reduces to (6) when $\lambda(t,x) \equiv \lambda$, i.e., the APU.

Under some proper assumptions (e.g., in El Karoui et al. 1997), $J^{\boldsymbol{\pi}}$, which is an $\mathcal{F}_t$-adapted process, solves the following backward stochastic differential equation (BSDE):

$$\mathrm{d}J_t^{\boldsymbol{\pi}} = -\lambda_t \mathcal{H}(\boldsymbol{\pi}_t)\left[(1-\gamma)J_t^{\boldsymbol{\pi}} + 1\right]\mathrm{d}t + \mathbf{Z}_t^{\boldsymbol{\pi}} \cdot \mathrm{d}\mathbf{B}_t, \;\; J_T^{\boldsymbol{\pi}} = U(W_T^{\boldsymbol{\pi}}), \tag{8}$$

where (and henceforth) $\lambda_t \equiv \lambda(t, X_t)$ and the notation $\cdot$ denotes the inner product. As explained in the Introduction, our formulation is motivated by the notion of "recursive utility" in the economics literature (Epstein and Zin, 1989; Duffie and Epstein, 1992), with consumption being the counterpart of randomization. In particular, the negative "dt" term in (8), as a function of $\boldsymbol{\pi}_t$ and $J_t^{\boldsymbol{\pi}}$, can be written as $f(\boldsymbol{\pi}, J, \lambda) = \lambda\mathcal{H}(\boldsymbol{\pi}) - \left[-\lambda(1-\gamma)\mathcal{H}(\boldsymbol{\pi})\right]J$. This term is the "*aggregator*" in the recursive utility jargon, while the term $-\lambda(1-\gamma)\mathcal{H}(\boldsymbol{\pi})$ corresponds to the discount rate that depreciates the future utility J into today's value.

We can formalize the above discussion to show a time preference on randomization implied by the definition of $J^{\boldsymbol{\pi}}$, in the same way as the Uzawa utility (Uzawa, 1968) for consumption. Indeed, we explicitly solve (7) as

$$J_t^{\boldsymbol{\pi}} = \mathbb{E}\left[\int_t^T e^{-\int_t^s -\lambda_\tau(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}\lambda_s\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + e^{-\int_t^T -\lambda_\tau(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}U(W_T^{\boldsymbol{\pi}})\middle|\mathcal{F}_t\right]. \tag{9}$$

This expression implies that the current weight on randomization depends on past randomization. Specifically, the (endogenous) temperature parameter for randomization now (compared to (6)) becomes $\lambda_s e^{-\int_0^s -\lambda_\tau(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}$. In other words, at any time s, a discount $e^{-\int_0^s -\lambda_\tau(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}$ is applied to (6). Moreover, empirical studies indicate that the typical risk-aversion parameter $\gamma > 1$ (see, e.g., Kydland and Prescott 1982), rendering $-\lambda_\tau(1-\gamma) > 0$. This implies that the more the investor has randomized in the past, the less weight she places on current randomization, which is intuitive and sensible. Next, consider a small risk aversion $\gamma \in (0,1)$, where the discounting factor becomes negative, which seems to incentivize larger randomization. However, there is a similar discount applied to the bequest utility. So, as randomization increases, the constant part of the bequest utility "$-\frac{1}{1-\gamma}$", which now is a *negative* number, will go even more negative. This introduces a proper trade-off between randomization and bequest utility even when risk aversion is weak, avoiding the ill-posedness that occurs in the APU case.

Technically, the modified objective function (9) makes our model mathematically tractable, as will be shown in the subsequent analysis.

Henceforth denote $\mu_t \equiv \mu(t, X_t)$ and $\sigma_t \equiv \sigma(t, X_t)$. We are now ready to formulate our RPU Merton problem, by first formally introducing the set of admissible controls.

Definition 1. An $\mathcal{F}_t$-adapted, probability-density-valued process $\boldsymbol{\pi} = \{\boldsymbol{\pi}_s, 0 \leqslant s \leqslant T\}$ is called an (open-loop) admissible control, if

(i) for each $0 \leqslant s \leqslant T$, $\boldsymbol{\pi}_s \in \mathcal{P}(\mathbb{R})$ a.s., where $\mathcal{P}(\mathbb{R})$ is the set of all probability densities on real numbers;

(ii) $\mathbb{E}\left[\int_0^T |\sigma_s|^2(\mathrm{Mean}(\boldsymbol{\pi}_s)^2 + \mathrm{Var}(\boldsymbol{\pi}_s))\mathrm{d}s\right] + \mathbb{E}\left[\int_0^T |\mu_s \,\mathrm{Mean}(\boldsymbol{\pi}_s)|\mathrm{d}s\right] < \infty$;

(iii) $\mathbb{E}\left[e^{\int_0^T 2\lambda_s |1-\gamma||\mathcal{H}(\boldsymbol{\pi}_s)|ds}\right] + \mathbb{E}\left[|U(W_T^{\boldsymbol{\pi}})|^2\right] < \infty$, where $\{X_s\}_{s\in[0,T]}$ and $\{W_s^{\boldsymbol{\pi}}\}_{s\in[0,T]}$ satisfy (2) and (5), respectively.

Given an admissible control $\boldsymbol{\pi}$ and an initial state pair (w_0, x_0), we define the recursive utility $J_t^{\boldsymbol{\pi}}$ through (7), where $\{X_t\}_{t\in[0,T]}$ and $\{W_t^{\boldsymbol{\pi}}\}_{t\in[0,T]}$ solve (2) and (5), respectively. A technical question with the above definition is whether the entropy term in (7) has a positive weight, which is answered by the following proposition. Hence, our RPU indeed incentivizes randomization.

Proposition 2. *For any admissible control $\boldsymbol{\pi}$, we have $(1-\gamma)J_t^{\boldsymbol{\pi}} + 1 > 0$ almost surely.*

Proof. Recall that RPU satisfies the BSDE (8). Given an admissible control $\boldsymbol{\pi}$, the condition (*ii*) in Definition 1 guarantees that the solution to (5) exists, and can be written as

$$W_t^{\boldsymbol{\pi}} = w_0 \exp\left\{ \int_0^t \left[r + (\mu_s - r)\,\mathrm{Mean}(\boldsymbol{\pi}_s) + \frac{1}{2}\sigma_s^2 \left(\mathrm{Mean}(\boldsymbol{\pi}_s)^2 + \mathrm{Var}(\boldsymbol{\pi}_s)\right)\right] \mathrm{d}s \right.$$
$$\left. + \sigma_s \left[\mathrm{Mean}(\boldsymbol{\pi}_s)\mathrm{d}B_s + \sqrt{\mathrm{Var}(\boldsymbol{\pi}_s)}\mathrm{d}\bar{B}_s\right] \right\} > 0.$$

In addition, (8) is a linear BSDE, and the condition (*iii*) in Definition 1 ensures that $U(W_T^{\boldsymbol{\pi}})$ and $\{\lambda(t,X_t)\mathcal{H}(\boldsymbol{\pi}_t)\}_{t\in[0,T]}$ are respectively square-integrable random variable and process. Hence by El Karoui et al. (1997, Proposition 2.2), (8) admits a unique square-integrable solution.

Define $\tilde{Y}_t^{\boldsymbol{\pi}} := (1-\gamma)J_t^{\boldsymbol{\pi}} + 1$. Applying Itô's lemma, we obtain that $\tilde{Y}^{\boldsymbol{\pi}}$ solves the following BSDE:

$$\mathrm{d}\tilde{Y}_t^{\boldsymbol{\pi}} = -(1-\gamma)\lambda(t,X_t)\mathcal{H}(\boldsymbol{\pi}_t)\tilde{Y}_t^{\boldsymbol{\pi}}\mathrm{d}t + (1-\gamma)\tilde{\mathbf{Z}}_t^{\boldsymbol{\pi}}\cdot\mathrm{d}\mathbf{B}_t,\ \tilde{Y}_T^{\boldsymbol{\pi}} = (W_T^{\boldsymbol{\pi}})^{1-\gamma} > 0.$$

From the comparison principle of linear BSDEs (see El Karoui et al. 1997, Corollary 2.2), it follows that $\tilde{Y}_t^{\boldsymbol{\pi}} > 0$ almost surely. Q.E.D.

To apply dynamic programming to problem (7), we further restrict our attention to feedback policies. A (randomized) *feedback policy* (or simply a *policy*) $\boldsymbol{\pi} = \boldsymbol{\pi}(\cdot,\cdot,\cdot)$ is a density-valued function of time and state, under which (2) and (5) become a Markovian system. For any initial time t and initial state (w,x), a policy $\boldsymbol{\pi}$ induces the open-loop control $\boldsymbol{\pi}_s = \boldsymbol{\pi}(s,W_s^{\boldsymbol{\pi}},X_s)$, where $\{X_s\}_{s\in[t,T]}$ and $\{W_s^{\boldsymbol{\pi}}\}_{s\in[t,T]}$ are the solutions to the corresponding Markovian system given $W_t^{\boldsymbol{\pi}} = w$ and $X_t = x$. Denote by Π the set of policies that induce admissible open-loop controls.

Given $\boldsymbol{\pi}\in\Pi$, define its *value function* $V^{\boldsymbol{\pi}}(\cdot,\cdot,\cdot)$ as

$$V^{\boldsymbol{\pi}}(t,w,x) := \mathbb{E}\Big[\int_t^T e^{-\int_t^s -\lambda(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}\lambda\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + e^{-\int_t^T -\lambda(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_\tau)\mathrm{d}\tau}U(W_T^{\boldsymbol{\pi}})\Big| W_t^{\boldsymbol{\pi}} = w, X_t = x\Big],$$
$$(t,w,x)\in[0,T]\times\mathbb{R}_+\times\mathbb{R}. \tag{10}$$

The Feynman-Kac formula yields that this function satisfies the PDE:

$$\begin{aligned}\frac{\partial V^{\boldsymbol{\pi}}(t,w,x)}{\partial t} &+ \Big(r + \big(\mu(t,x)-r\big)\,\mathrm{Mean}\,(\boldsymbol{\pi}(t,w,x))\Big)wV_w^{\boldsymbol{\pi}}(t,w,x)\\ &+ \frac{1}{2}\sigma^2(t,x)\Big(\mathrm{Mean}\,(\boldsymbol{\pi}(t,w,x))^2 + \mathrm{Var}\,(\boldsymbol{\pi}(t,w,x))\Big)w^2V_{ww}^{\boldsymbol{\pi}}(t,w,x)\\ &+ m(t,x)V_x^{\boldsymbol{\pi}}(t,w,x) + \frac{1}{2}\nu^2(t,x)V_{xx}^{\boldsymbol{\pi}}(t,w,x) + \rho\nu(t,x)\sigma(t,x)\,\mathrm{Mean}(\boldsymbol{\pi})wV_{xw}^{\boldsymbol{\pi}}(t,w,x)\\ &+ \lambda(t,x)\mathcal{H}\,(\boldsymbol{\pi}(t,w,x))\big[(1-\gamma)V^{\boldsymbol{\pi}}(t,w,x)+1\big] = 0,\ V^{\boldsymbol{\pi}}(T,w,x) = U(w).\end{aligned} \tag{11}$$

Using the relation between BSDEs and PDEs, we can represent the recursive utility $J^{\boldsymbol{\pi}}$ via the value function $V^{\boldsymbol{\pi}}$:

$$J_t^{\boldsymbol{\pi}} = V^{\boldsymbol{\pi}}(t,W_t^{\boldsymbol{\pi}},X_t),\ \text{a.s.},\quad t\in[0,T], \tag{12}$$

where $\{X_t\}_{t\in[0,T]}$ and $\{W_t^{\boldsymbol{\pi}}\}_{t\in[0,T]}$ satisfy (2) and (5) with $W_0^{\boldsymbol{\pi}}=w_0$ and $X_0=x_0$, respectively.

Finally, we define the *optimal* value function as

$$V(t,w,x):=\sup_{\boldsymbol{\pi}\in\Pi} J^{\boldsymbol{\pi}}(t,w,x),\ \ (t,w,x)\in[0,T]\times\mathbb{R}_+\times\mathbb{R}. \tag{13}$$

3. Theoretical Analysis

3.1. Gaussian Randomization

It is straightforward, as in Wang et al. (2020), to derive that the optimal value function V satisfies the following HJB equation via dynamic programming:

$$\begin{aligned}&\frac{\partial V}{\partial t}+\sup_{\boldsymbol{\pi}\in\mathcal{P}(\mathbb{R})}\Bigg\{\Big(r+\big(\mu(t,x)-r\big)\operatorname{Mean}(\boldsymbol{\pi})\Big)wV_w+\frac{1}{2}\sigma^2(t,x)\Big(\operatorname{Mean}(\boldsymbol{\pi})^2+\operatorname{Var}(\boldsymbol{\pi})\Big)w^2V_{ww}\\&+m(t,x)V_x+\frac{1}{2}\nu^2(t,x)V_{xx}+\rho\nu(t,x)\sigma(t,x)\operatorname{Mean}(\boldsymbol{\pi})wV_{wx}+\lambda(t,x)\mathcal{H}(\boldsymbol{\pi})\big[(1-\gamma)V+1\big]\Bigg\}=0,\end{aligned} \tag{14}$$

with the terminal condition $V(T,w,x)=U(w)=\frac{w^{1-\gamma}-1}{1-\gamma}$.

At first glance, equation (14) is a highly nonlinear PDE and appears hard to analyze. However, we can reduce it to a simpler PDE based on which the optimal randomized policy can be explicitly represented.

Theorem 1. *Suppose u is a classical solution of the following PDE*

$$\begin{aligned}&\frac{\partial u}{\partial t}+(1-\gamma)r+m(t,x)u_x+\frac{1}{2}\nu^2(t,x)(u_{xx}+u_x^2)+\frac{(1-\gamma)\lambda(t,x)}{2}\log\frac{2\pi\lambda(t,x)}{\gamma\sigma^2(t,x)}\\&+\frac{1-\gamma}{2\gamma}\left[\frac{\big(\mu(t,x)-r\big)^2}{\sigma^2(t,x)}+\frac{2\rho\big(\mu(t,x)-r\big)\nu(t,x)}{\sigma(t,x)}u_x+\rho^2\nu^2(t,x)u_x^2\right]=0,\end{aligned} \tag{15}$$

with the terminal condition $u(T,x)=0$, and the derivatives of u up to the second order have polynomial growth. Let $\boldsymbol{\pi}^*(t,x)$ be a normal distribution with

$$\operatorname{Mean}(\boldsymbol{\pi}^*(t,x))=\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)}+\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}u_x(t,x),\ \operatorname{Var}(\boldsymbol{\pi}^*(t,x))=\frac{\lambda(t,x)}{\gamma\sigma^2(t,x)}. \tag{16}$$

If $\boldsymbol{\pi}^*\in\Pi$, then it is the optimal randomized policy. Furthermore, the optimal value function is $V(t,w,x)=\frac{w^{1-\gamma}e^{u(t,x)}-1}{1-\gamma}$.

Proof. First, let us show how to deduce (16) from (14). Note that the "supermum" in (14) can be achieved via a two-stage optimization procedure: first maximize over distributions with a *fixed* mean and variance pair, and then maximize over all possible such pairs. For the first problem, the entropy is maximized at a normal distribution with the fixed mean and variance, with the maximum entropy value $\mathcal{H}(\boldsymbol{\pi}) = \frac{1}{2}\log\big(2\pi e\,\mathrm{Var}(\boldsymbol{\pi})\big)$ where $\mathrm{Var}(\boldsymbol{\pi})$ is the given fixed variance. Therefore, (14) can be simplified as

$$\frac{\partial V}{\partial t} + \sup_{\mathrm{Mean}(\boldsymbol{\pi}),\mathrm{Var}(\boldsymbol{\pi})}\Bigg\{\Big(r+\big(\mu(t,x)-r\big)\,\mathrm{Mean}(\boldsymbol{\pi})\Big)wV_w + \frac{1}{2}\sigma^2(t,x)\Big(\mathrm{Mean}(\boldsymbol{\pi})^2 + \mathrm{Var}(\boldsymbol{\pi})\Big)w^2V_{ww}$$
$$+ m(t,x)V_x + \frac{1}{2}\nu^2(t,x)V_{xx} + \rho\nu(t,x)\sigma(t,x)\,\mathrm{Mean}(\boldsymbol{\pi})wV_{wx} + \frac{\lambda(t,x)}{2}\log\big(2\pi e\,\mathrm{Var}(\boldsymbol{\pi})\big)\big[(1-\gamma)V+1\big]\Bigg\} = 0. \tag{17}$$

To analyze the above PDE, we start with the ansatz that $V(t,w,x) = \frac{w^{1-\gamma}}{1-\gamma}v(t,x) - \frac{1}{1-\gamma}$ for some function v. Then (17) becomes

$$\frac{w^{1-\gamma}}{1-\gamma}\frac{\partial v}{\partial t} + \sup_{\mathrm{Mean}(\boldsymbol{\pi}),\mathrm{Var}(\boldsymbol{\pi})}\Bigg\{\Big(r+\big(\mu(t,x)-r\big)\,\mathrm{Mean}(\boldsymbol{\pi})\Big)w^{1-\gamma}v - \frac{\gamma}{2}\sigma^2(t,x)\Big(\mathrm{Mean}(\boldsymbol{\pi})^2 + \mathrm{Var}(\boldsymbol{\pi})\Big)w^{1-\gamma}v$$
$$+ m(t,x)\frac{w^{1-\gamma}}{1-\gamma}v_x + \frac{1}{2}\nu^2(t,x)\frac{w^{1-\gamma}}{1-\gamma}v_{xx} + \rho\nu(t,x)\sigma(t,x)\,\mathrm{Mean}(\boldsymbol{\pi})w^{1-\gamma}v_x + \frac{\lambda(t,x)}{2}\log\big(2\pi e\,\mathrm{Var}(\boldsymbol{\pi})\big)w^{1-\gamma}v\Bigg\} = 0. \tag{18}$$

If $v > 0$ (to be verified later), then the first-order conditions of the maximization problem on the left-hand side of the above equation yield the maximizers

$$\mathrm{Mean}(\boldsymbol{\pi}^*) = \frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)v_x(t,x)}{\gamma\sigma(t,x)v(t,x)}, \quad \mathrm{Var}(\boldsymbol{\pi}^*) = \frac{\lambda(t,x)}{\gamma\sigma^2(t,x)}. \tag{19}$$

Plugging the above into (18) and canceling the term $w^{1-\gamma}$, we have that v satisfies the following nonlinear PDE:

$$\begin{aligned}&\frac{\partial v}{\partial t} + (1-\gamma)rv + m(t,x)v_x + \frac{1}{2}\nu^2(t,x)v_{xx} + \frac{(1-\gamma)\lambda(t,x)v}{2}\log\frac{2\pi\lambda(t,x)}{\gamma\sigma^2(t,x)}\\ &\quad + \frac{1-\gamma}{2\gamma}\Bigg[\frac{\big(\mu(t,x)-r\big)^2}{\sigma^2(t,x)}v + \frac{2\rho\big(\mu(t,x)-r\big)\nu(t,x)}{\sigma(t,x)}v_x + \rho^2\nu^2(t,x)\frac{v_x^2}{v}\Bigg] = 0,\end{aligned} \tag{20}$$

with the terminal condition $v(T,x) = 1$.

Letting u be the solution to (15), we can directly verify that $v(t,x) = e^{u(t,x)} > 0$ solves (20). The desired expression (16) now follows from (19).

Next we prove that the policy with (16) is optimal and $V(t,w,x)=\frac{w^{1-\gamma}e^{u(t,x)}-1}{1-\gamma}$ is the optimal value function. Denote $K_t^{\boldsymbol{\pi}}=\int_0^t\lambda(s,X_s)(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s$. Apply Itô's formula to $e^{K_t}V(t,W_t^{\boldsymbol{\pi}},X_t)$ to get

$$
\begin{aligned}
&\mathrm{d}\left(e^{K_t^{\boldsymbol{\pi}}}V(t,W_t^{\boldsymbol{\pi}},X_t)\right)\\
=&\Bigg\{\frac{\partial V}{\partial t}+\left[r+(\mu(t,X_t)-r)\operatorname{Mean}(\boldsymbol{\pi}_t)\right]W_t^{\boldsymbol{\pi}}V_w+m(t,X_t)V_x+\frac{1}{2}\sigma^2(t,X_t)\left[\operatorname{Mean}(\boldsymbol{\pi}_t)^2+\operatorname{Var}(\boldsymbol{\pi}_t)\right](W_t^{\boldsymbol{\pi}})^2V_{ww}\\
&+\frac{1}{2}\nu^2(t,X_t)V_{xx}+\rho\nu(t,X_t)\sigma(t,X_t)\operatorname{Mean}(\boldsymbol{\pi}_t)W_t^{\boldsymbol{\pi}}V_w+\lambda(1-\gamma)\mathcal{H}(\boldsymbol{\pi}_t)V\Bigg\}e^{K_t^{\boldsymbol{\pi}}}\mathrm{d}t\\
&+\Bigg\{\left[\sigma(t,X_t)\operatorname{Mean}(\boldsymbol{\pi}_t)W_t^{\boldsymbol{\pi}}V_w+\rho\nu(t,X_t)V_x\right]\mathrm{d}B_t\\
&+\sqrt{1-\rho^2}\nu(t,X_t)V_x\mathrm{d}\tilde{B}_t+\sqrt{\operatorname{Var}(\boldsymbol{\pi}_t)}\sigma(t,X_t)W_t^{\boldsymbol{\pi}}V_w\mathrm{d}\bar{B}_t\Bigg\}e^{K_t^{\boldsymbol{\pi}}}.
\end{aligned}
$$

Define a sequence of stopping times $\tau_n=\inf\{t\geqslant 0:|X_t|\vee(W_t^{\boldsymbol{\pi}})^{1-\gamma}\vee e^{K_t^{\boldsymbol{\pi}}}\geqslant n\}$. Then by (17), we have

$$
\begin{aligned}
&e^{K_{T\wedge\tau_n}^{\boldsymbol{\pi}}}V(T\wedge\tau_n,W_{T\wedge\tau_n}^{\boldsymbol{\pi}},X_{T\wedge\tau_n})-V(0,W_0^{\boldsymbol{\pi}},X_0)\\
\leqslant&\int_0^{T\wedge\tau_n}-\lambda(s,X_s)\mathcal{H}(\boldsymbol{\pi}_s)e^{K_s^{\boldsymbol{\pi}}}\mathrm{d}s+e^{K_s^{\boldsymbol{\pi}}}(W_s^{\boldsymbol{\pi}})^{1-\gamma}e^{u(s,X_s)}\Bigg\{\left[\sigma(s,X_s)\operatorname{Mean}(\boldsymbol{\pi}_s)+\rho\nu(s,X_s)\frac{u_x(s,X_s)}{1-\gamma}\right]\mathrm{d}B_s\\
&+\sqrt{1-\rho^2}\nu(s,X_s)\frac{u_x(s,X_s)}{1-\gamma}\mathrm{d}\tilde{B}_s+\sqrt{\operatorname{Var}(\boldsymbol{\pi}_s)}\sigma(s,X_s)\mathrm{d}\bar{B}_s\Bigg\},
\end{aligned}
$$

while "=" holds when $\boldsymbol{\pi}=\boldsymbol{\pi}^*$ as given in (16).

Recall that u and its derivatives have polynomial growth in x; hence we have an estimate about the quadratic variation term above

$$
\begin{aligned}
&\mathbb{E}\Bigg[\int_0^{T\wedge\tau_n}e^{2K_s^{\boldsymbol{\pi}}}(W_s^{\boldsymbol{\pi}})^{2(1-\gamma)}e^{2u(s,X_s)}\Bigg\{\left[\sigma(s,X_s)\operatorname{Mean}(\boldsymbol{\pi}_s)+\rho\nu(s,X_s)\frac{u_x(s,X_s)}{1-\gamma}\right]^2\\
&+(1-\rho^2)\nu^2(s,X_s)\frac{u_x^2(s,X_s)}{(1-\gamma)^2}+\operatorname{Var}(\boldsymbol{\pi}_s)\sigma^2(s,X_s)\Bigg\}\mathrm{d}s\Bigg]\\
\leqslant&\mathbb{E}\left[\int_0^T C_n\left\{\sigma^2(s,X_s)\left[\operatorname{Mean}(\boldsymbol{\pi}_s)^2+\operatorname{Var}(\boldsymbol{\pi}_s)\right]+1\right\}\mathrm{d}s\right]
\end{aligned}
$$

for some constants $C_n > 0$. Therefore, the expectations of the corresponding stochastic integrals are all 0. When $\boldsymbol{\pi}$ is admissible, by the condition (ii) in Definition 1, we deduce

$$
\begin{aligned}
&V(0,w_0,x_0)\\
\geqslant&\mathbb{E}\left[e^{K^{\boldsymbol{\pi}}_{T\wedge\tau_n}}V(T\wedge\tau_n,W^{\boldsymbol{\pi}}_{T\wedge\tau_n},X_{T\wedge\tau_n})+\int_0^{T\wedge\tau_n}\lambda(s,X_s)\mathcal{H}(\boldsymbol{\pi}_s)e^{K^{\boldsymbol{\pi}}_s}\mathrm{d}s\right]\\
=&\mathbb{E}\left[e^{K^{\boldsymbol{\pi}}_{T\wedge\tau_n}}\frac{(W^{\boldsymbol{\pi}}_{T\wedge\tau_n})^{1-\gamma}}{1-\gamma}e^{u(T\wedge\tau_n,X_{T\wedge\tau_n})}-\frac{e^{K^{\boldsymbol{\pi}}_{T\wedge\tau_n}}}{1-\gamma}+\int_0^{T\wedge\tau_n}\lambda(s,X_s)\mathcal{H}(\boldsymbol{\pi}_s)e^{K^{\boldsymbol{\pi}}_s}\mathrm{d}s\right]\\
=&\mathbb{E}\left[e^{K^{\boldsymbol{\pi}}_T}\frac{(W^{\boldsymbol{\pi}}_T)^{1-\gamma}}{1-\gamma}\mathbb{1}_{\{\tau_n>T\}}+e^{K^{\boldsymbol{\pi}}_{\tau_n}}\frac{(W^{\boldsymbol{\pi}}_{\tau_n})^{1-\gamma}}{1-\gamma}e^{u(\tau_n,X_{\tau_n})}\mathbb{1}_{\{\tau_n\leqslant T\}}-\frac{e^{K^{\boldsymbol{\pi}}_{T\wedge\tau_n}}}{1-\gamma}+\int_0^{T\wedge\tau_n}\lambda(s,X_s)\mathcal{H}(\boldsymbol{\pi}_s)e^{K^{\boldsymbol{\pi}}_s}\mathrm{d}s\right].
\end{aligned}
$$

By the condition (iii) in Definition 1, we have

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^T\lambda(t,X_t)|\mathcal{H}(\boldsymbol{\pi}_t)|e^{K^{\boldsymbol{\pi}}_t}\mathrm{d}t\right]\leqslant\mathbb{E}\left[\int_0^T\lambda(t,X_t)|\mathcal{H}(\boldsymbol{\pi}_t)|e^{\int_0^t\lambda|1-\gamma||\mathcal{H}(\boldsymbol{\pi}_s)|\mathrm{d}s}\mathrm{d}t\right]\\
\leqslant&\mathbb{E}\left[\int_0^T\lambda(t,X_t)|\mathcal{H}(\boldsymbol{\pi}_t)|\mathrm{d}te^{\int_0^T\lambda|1-\gamma||\mathcal{H}(\boldsymbol{\pi}_s)|\mathrm{d}s}\right]\\
\leqslant&\left(\mathbb{E}\left[\left(\int_0^T\lambda(t,X_t)|\mathcal{H}(\boldsymbol{\pi}_t)|\mathrm{d}t\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[e^{\int_0^T2\lambda|1-\gamma||\mathcal{H}(\boldsymbol{\pi}_s)|\mathrm{d}s}\right]\right)^{1/2}<\infty,
\end{aligned}
$$

and

$$
\mathbb{E}\left[e^{K^{\boldsymbol{\pi}}_T}(W^{\boldsymbol{\pi}}_T)^{1-\gamma}\right]\leqslant\left(\mathbb{E}\left[e^{\int_0^T2\lambda|1-\gamma||\mathcal{H}(\boldsymbol{\pi}_s)|\mathrm{d}s}\right]\right)^{1/2}\left(\mathbb{E}\left[(W^{\boldsymbol{\pi}}_T)^{2-2\gamma}\right]\right)^{1/2}<\infty.
$$

Taking $\limsup_{n\to\infty}$, we conclude from the dominated convergence theorem and Fatou's lemma that

$$
V(0,w_0,x_0)\geqslant\mathbb{E}\left[e^{K^{\boldsymbol{\pi}}_T}U(W^{\boldsymbol{\pi}}_T)+\int_0^T\lambda(s,X_s)\mathcal{H}(\boldsymbol{\pi}_s)e^{K^{\boldsymbol{\pi}}_s}\mathrm{d}s\right]=J^{\boldsymbol{\pi}}_0.
$$

To show "=" for $\boldsymbol{\pi}^*$, we can directly verify that $V_t = V(t, W^{\boldsymbol{\pi}^*}_t, X_t)$ satisfies the BSDE (8), and hence V_t is the recursive utility process under $\boldsymbol{\pi}^*$. This completes our proof. Q.E.D.

Some remarks are in order. First, the optimal randomized policy follows *Gaussian* with its mean and variance explicitly given via the PDE (15), even if the current setup is neither LQ nor mean-variance. The policy variance depends only on the exogenously given primary temperature function λ, the risk aversion parameter γ, and the instantaneous variance function σ^2. A larger primary temperature increases the level of randomization, while a greater risk aversion or volatility reduces it. These results are mathematically consistent with those in Wang et al. (2020) and Dai et al. (2023), even though in a different context (i.e., RL), who consider an LQ control and an equilibrium mean-variance criterion, respectively. Second, the mean of the optimal policy consists of two parts: a myopic part $\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)}$ independent of randomization, and a hedging part represented by

$$\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}u_x(t,x) = \frac{\mathrm{Cov}(\mathrm{d}X_t, \mathrm{d}\log S_t)}{\mathrm{Var}(\mathrm{d}\log S_t)}u_x(t,X_t).$$

Note that hedging is needed due to the presence of the factor X even in the classical Merton setting; yet the level of hedging is *altered* by the agent randomization because u depends on the choice of λ via the PDE (15). As a result, unlike the previous works (e.g., Wang et al. 2020, Wang and Zhou 2020, and Dai et al. 2023) where the optimal policies depend on randomization only through variance and are thus unbiased, the optimal policies here are generally *biased* and the degree of biasedness depends on that of randomization. We will investigate this feature in more detail in the following subsections. Finally, the resulting weight of randomization in the objective function is $\lambda(t,x)[(1-\gamma)V(t,w,x)+1] = \lambda(t,x)w^{1-\gamma}e^{u(t,x)} > 0$. When $\lambda(t,x)$ is a constant and $\gamma=1$, $u\equiv 0$ is the solution to (15) and, as a consequence, the weight reduces to a constant as in the APU objective (6).

3.2. When is the Optimal Randomized Policy Unbiased?

When $\lambda\equiv 0$, the optimal Gaussian distribution in Theorem 1 degenerates into the Dirac measure concentrating on the mean, $\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}u_x(t,x)$, where u solves (15) with $\lambda\equiv 0$. This is Merton's strategy for the classical problem (in the incomplete market) without preference for randomization. We refer to the case $\lambda\equiv 0$ as the "classical case" in the rest of this paper. As we have pointed out, unlike the existing results, the mean of the optimal randomized policy does not generally coincide with that of the classical counterpart due to an interaction between randomization and hedging.[12]

There are, however, special circumstances even in our setting where the optimal Gaussian policy becomes unbiased. According to (16), the part that causes biases is the hedging demand, $\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}u_x(t,x)$. Hence, if $\nu\equiv 0$ or $\rho=0$ (i.e., the factor X is deterministic or evolves independently from the stock price), then this part vanishes and the optimal policy becomes unbiased. In these cases, changes in the market factor do not affect the stock return or there is no hedging need against the factor, and thus a myopic policy irrelevant to our choice of λ is dynamically optimal. Next, note that the only difference between the classical and RPU problems is reflected by the extra term $\frac{(1-\gamma)\lambda(t,x)}{2}\log\frac{2\pi\lambda(t,x)}{\gamma\sigma^2(t,x)}$ in the PDE (15). If $\gamma=1$ (log-utility), then, for any choice of the function λ, this extra term vanishes, and hence the unbiasedness holds. More generally, if one chooses λ such that $\lambda(t,x)\log\frac{2\pi\lambda(t,x)}{\gamma\sigma^2(t,x)}$ is independent of x, then, taking derivative in x on both sides of (15) yields that u_x satisfies the same PDE regardless of whether $\lambda=0$ or not. This in turn implies that the hedging part $\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}u_x(t,x)$ is independent of λ, leading to the unbiasedness of the optimal Gaussian policy.

A discussion of the unbiasedness will also be given from the BSDE perspective in Appendix C.

3.3. An Asymptotic Analysis on λ

The preference for randomization induces a different objective function for the investor and in general makes the optimal policy biased. In particular, randomization increases the uncertainty of the wealth process and thus lowers the bequest utility. It is interesting to investigate the financial losses due to this preference for randomization and quantify the impact of such a bias. We carry out this investigation by an asymptotic analysis on the PDE (15) in the small parameter λ, which is henceforth assumed to be a *constant* (instead of a function) $\lambda(t,x)\equiv\lambda$, leading to asymptotic expansions of the optimal policy along with its value function.

We denote by $u^{(0)}$ the solution to (15) with $\lambda=0$ and by $V^{(0)}$ the optimal value function for the classical problem (i.e., when $\lambda=0$). It follows from Theorem 1 that $V^{(0)}(t,w,x)=\frac{w^{1-\gamma}e^{u^{(0)}(t,x)}-1}{1-\gamma}$. For any $\lambda>0$, let $\boldsymbol{\pi}^*$ be the optimal randomized policy, and $V^{(\lambda)}$ be the value function of the original *non-randomized* problem under the deterministic policy that is the mean of $\boldsymbol{\pi}^*$:

$$V^{(\lambda)}(t,w,x)=\mathbb{E}\left[U\left(W_T^{\widehat{\boldsymbol{\pi}^*}}\right)\middle|W_t^{\widehat{\boldsymbol{\pi}^*}}=w,X_t=x\right], \tag{21}$$

where $\widehat{\boldsymbol{\pi}^*}(t,x)=\mathcal{N}\left(\mathrm{Mean}\left(\boldsymbol{\pi}^*(t,x)\right),0\right)$.

The results of our asymptotic analysis involve several functions, among which $u^{(1)}$ and $u^{(2)}$ are respectively the solutions of the following PDEs:

$$\begin{aligned}&\frac{\partial u^{(1)}}{\partial t}+m(t,x)u_x^{(1)}+\frac{1}{2}\nu^2(t,x)u_{xx}^{(1)}+\frac{1-\gamma}{2}\log\frac{2\pi}{\gamma\sigma^2(t,x)}\\&+(1-\gamma)\rho\nu(t,x)\sigma(t,x)\frac{\mu(t,x)-r+\rho\nu(t,x)\sigma(t,x)u_x^{(0)}}{\gamma\sigma^2(t,x)}u_x^{(1)}=0,\ \ u^{(1)}(T,x)=0,\end{aligned}\tag{22}$$

and

$$\begin{aligned}&\frac{\partial u^{(2)}}{\partial t}+m(t,x)u_x^{(2)}+\frac{1}{2}\nu^2(t,x)u_{xx}^{(2)}+\frac{1-\gamma}{2\gamma}\rho^2\nu^2(t,x){u_x^{(1)}}^2\\&+(1-\gamma)\rho\nu(t,x)\sigma(t,x)\frac{\mu(t,x)-r+\rho\nu(t,x)\sigma(t,x)u_x^{(0)}}{\gamma\sigma^2(t,x)}u_x^{(2)}=0,\ \ u^{(2)}(T,x)=0.\end{aligned}\tag{23}$$

Moreover, $\phi^{(2)}$ is the solution to the PDE:

$$\begin{aligned}&\frac{\partial\phi^{(2)}}{\partial t}+m(t,x)\phi_x^{(2)}+\frac{1}{2}\nu^2(t,x)(\phi_{xx}^{(2)}+2u_x^{(0)}\phi_x^{(2)})\\&\quad+\frac{(1-\gamma)\rho\nu(t,x)}{\gamma\sigma(t,x)}\big[\mu(t,x)-r+\rho\nu(t,x)\sigma(t,x)u_x^{(0)}\big]\phi_x^{(2)}\\&\quad-\frac{(1-\gamma)\rho^2\nu^2(t,x)}{2\gamma}{u_x^{(1)}}^2=0,\ \ \phi^{(2)}(T,x)=0.\end{aligned}\tag{24}$$

Lemma 1. *The solution u of* (15) *admits the following Taylor expansion with respect to λ:*

$$u(t,x)=u^{(0)}(t,x)+\frac{(1-\gamma)(T-t)}{2}\lambda\log\lambda+\lambda u^{(1)}(t,x)+\lambda^2u^{(2)}(t,x)+O(\lambda^3).\tag{25}$$

Proof. The expansion (25) is motivated by observing that the λ terms in (15) are $\frac{1-\gamma}{2}\lambda\log\lambda$ plus $O(\lambda)$. Plugging u with the expansion into (1), using the equation of $u^{(0)}$ and equating the λ and λ^2 terms in the resulting equation, we easily derive the equations (22) and (23). Q.E.D.

The policy that coincides with the mean of the optimal randomized policy is sub-optimal for the classical Merton problem due to randomization, which can be considered as a loss in the initial wealth. For each fixed initial time-factor pair (t,x), we define the *equivalent relative wealth loss* to be $\Delta=\Delta(t,x)$ such that the investor is indifferent between obtaining the optimal value without the preference for randomization with initial endowment $w(1-\Delta)$ and getting the value of the mean policy, Mean($\boldsymbol{\pi}^*$), with initial endowment w. That is, Δ is such that $V^{(0)}\big(t,w(1-\Delta),x\big)=V^{(\lambda)}(t,w,x)$. The equivalent relative wealth loss Δ measures the relative cost that the agent is willing to pay for the pleasure of randomization.

The following theorem quantifies the bias of the optimal randomized policy, the relative loss of utility, and the equivalent relative wealth loss, all in terms of λ asymptotically.

THEOREM 2. *The asymptotic expansion of the mean of the optimal randomized policy is*

$$\mathrm{Mean}(\boldsymbol{\pi}^*(t,x)) = a^*(t,x) + \lambda \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)} u_x^{(1)}(t,x) + O(\lambda^2),$$

where a^* is the optimal policy for the classical case. Moreover, we have

$$V^{(\lambda)}(t,w,x) = \frac{w^{1-\gamma}}{1-\gamma} \exp\left(u^{(0)}(t,x) + \lambda^2 \phi^{(2)}(t,x) + O(\lambda^3)\right) - \frac{1}{1-\gamma},$$

along with the relative utility loss

$$\left|\frac{V^{(\lambda)}(t,w,x) - V^{(0)}(t,w,x)}{V^{(0)}(t,w,x)}\right| = \lambda^2 |\phi^{(2)}(t,x)| \cdot \left|1 + \frac{1}{(1-\gamma)V^{(0)}(t,w,x)}\right| + O(\lambda^3).$$

Finally, the equivalent relative wealth loss is

$$\Delta(t,x) = -\frac{\lambda^2 \phi^{(2)}(t,x)}{1-\gamma} + O(\lambda^3).$$

Proof. Theorem 1 along with Lemma 1 imply that the mean of the optimal randomized policy is expanded as

$$\begin{aligned} a^{(\lambda)}(t,x) :=& \frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}\left[u_x^{(0)}(t,x) + \lambda u_x^{(1)}(t,x) + \lambda^2 u_x^{(2)}(t,x) + O(\lambda^3)\right] \\ =& a^*(t,x) + \lambda \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)} u_x^{(1)}(t,x) + \lambda^2 \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)} u_x^{(2)}(t,x) + O(\lambda^3), \end{aligned}$$

where $a^*(t,x) = \frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)} u_x^{(0)}(t,x)$ is the optimal policy for the classical case.

Recall that $V^{(\lambda)}$ is the value function of the classical problem under the deterministic policy $a^{(\lambda)}$. By the Feynman-Kac formula, $V^{(\lambda)}$ satisfies the PDE:

$$\begin{aligned} V_t^{(\lambda)} +& \Big(r + \big(\mu(t,x) - r\big) a^{\lambda}(t,x)\Big) w V_w^{(\lambda)} + \frac{1}{2}\sigma^2(t,x) {a^{\lambda}}^2(t,x) w^2 V_{ww}^{(\lambda)} \\ &+ m(t,x) V_x^{(\lambda)} + \frac{1}{2}\nu^2(t,x) V_{xx}^{(\lambda)} + \rho\nu(t,x)\sigma(t,x) a^{(\lambda)}(t,x) w V_{wx}^{(\lambda)} = \beta V^{(\lambda)}, \end{aligned} \tag{26}$$

with the terminal condition $V^{(\lambda)}(T,w,x) = U(w) = \frac{w^{1-\gamma}-1}{1-\gamma}$. Conjecturing $V^{(\lambda)}(t,w,x) = \frac{w^{1-\gamma}-1}{1-\gamma} \exp\{\psi^{(\lambda)}(t,x)\} + \varphi^{(\lambda)}(t,x)$ and putting it to (26), we obtain $-\frac{1}{1-\gamma}\exp\{\psi^{(\lambda)}(t,x)\} + \varphi^{(\lambda)}(t,x) = -\frac{e^{-\beta(T-t)}}{1-\gamma}$, where $\psi^{(\lambda)}$ satisfies

$$\begin{aligned} \psi_t^{(\lambda)} &+ m(t,x)\psi_x^{(\lambda)} + \frac{1}{2}\nu^2(t,x)(\psi_{xx}^{(\lambda)} + {\psi_x^{(\lambda)}}^2) + (1-\gamma)\rho\nu(t,x)\sigma(t,x) a^{(\lambda)}(t,x)\psi_x^{(\lambda)} \\ &+ [(1-\gamma)r - \beta] + (1-\gamma)\Big[(\mu(t,x)-r)a^{(\lambda)}(t,x) - \frac{\gamma}{2}\sigma^2(t,x){a^{(\lambda)}}^2(t,x)\Big] = 0, \ \psi^{(\lambda)}(T,x) = 0. \end{aligned}$$

However, when $\lambda = 0$, $\psi^{(0)} = u^{(0)}$, which motivates us to expand $\psi^{(\lambda)} = u^{(0)} + \lambda\phi^{(1)} + \lambda^2\phi^{(2)} + O(\lambda^3)$. Substituting this to the above equation, we deduce that $\phi^{(1)}$ satisfies

$$\begin{aligned}&\phi_t^{(1)} + m(t,x)\phi_x^{(1)} + \frac{1}{2}\nu^2(t,x)(\phi_{xx}^{(1)} + 2u_x^{(0)}\phi_x^{(1)}) \\ &\quad + \frac{(1-\gamma)\rho\nu(t,x)}{\gamma\sigma(t,x)}\big[\mu(t,x) - r + \rho\nu(t,x)\sigma(t,x)u_x^{(0)}\big]\phi_x^{(1)} = 0,\ \phi^{(1)}(T,x) = 0.\end{aligned}$$

Because $\phi^{(1)} \equiv 0$ is a solution to this linear PDE, it can be easily checked that (24) is the equation satisfied by $\phi^{(2)}$. The proof is complete. Q.E.D.

Even though the policy bias is of order $O(\lambda)$, both the relative utility loss and the equivalent relative wealth loss are of order $O(\lambda^2)$. So financial and utility losses due to preference for randomization are of higher order of the policy deviation.

4. Conclusions

This paper aims to address the prevalent appetite for randomization in a dynamic setting of Merton's problem. We introduce the RPU with entropy functions to represent the preference for stochastic choices, and prove that the optimal policy is Gaussian in a general Markovian incomplete market with CRRA bequest utility. The mean of the Gaussian policy generally differs from the classical Merton solution due to intertemporal hedging demand. An asymptotic expansion in temperature quantifies the deviation of the optimal mean from the classical benchmark and the associated wealth loss as the financial cost of the preference for randomization.

This work opens the gate to several directions of interesting future research. For example, dynamic and recursive preference for randomization calls for more solid micro-economics underpinning. This includes, among others, extending the current framework to include consumption which would allow a reexamination of RPU, and conducting empirical analysis that could help quantify the preference for randomization and assess whether RPU can better explain asset pricing and consumption patterns. It is also interesting to investigate the model-free, RL setting of the current problem, where randomization is out of both necessity (for exploration) and pleasure (for additional utility). In such a setting, how to distinguish and disentangle these two and how do they interact?

Along a different line, for Barberis' model of optimal exit from casino gambling (Barberis, 2012) featuring Kahneman and Tversky's cumulative prospective theory (CPT), He et al. (2017) and Hu et al. (2023) show that allowing randomized strategies *strictly* improves the optimal CPT value and they attribute this to the non-concavity of the S-shaped utility function in CPT. This suggests that preference for randomization may be implicitly captured by certain non-concave preferences without having to add explicitly a perturbed utility. Technically, it would be valuable to inquire the impact of alternative perturbed functions such as the Tsallis or Rényi entropy and/or alterative bequest utility functions beyond CRRA/CARA, as well as to study RPU for general stochastic control problems.

Acknowledgments

Dai acknowledges the support of Hong Kong GRF (15213422 & 15217123), The Hong Kong Polytechnic University Research Grants (P0039114, P0042456, P0042708, and P0045342), and the National Natural Science Foundation of China (72432005). Dong is supported by the National Natural Science Foundation of China (12071333 & 12101458). Jia acknowledges the support of The Chinese University of Hong Kong start-up grant. Zhou is supported by a start-up grant and the Nie Center for Intelligent Asset Management at Columbia University. His work is also part of a Columbia-CityU/HK collaborative project that is supported by the InnoHK Initiative, The Government of the HKSAR, and the AIFT Lab.

Endnotes

1. Here "state" includes *observable* variables including portfolio worth, known factors and stock prices.

2. "Omakase", meaning in Japanese "to entrust", originally refers to a set sushi course whose contents are completely determined by the chef based on his/her observation of an individual customer. Therefore, the surprise from the revelation of each sushi dish is a major fun part of the dining experience. Nowadays, omakase meals popular in many countries are often pre-announced set menus, thereby losing their original feature (and excitement) of randomization.

3. Also called "mystery boxes", blind boxes are curated, randomized packages sold for a fixed price, containing undisclosed items ranging from electronics and collectibles to clothing and snack boxes. The ultra-popular Labubu figures produced by Pop Mart are such an example.

4. For finite choice sets, APU connects to discrete-choice and random-utility models (McFadden, 1974, 2001; Anderson et al., 1992; Berry, 1994; Machina, 1985; Mattsson and Weibull, 2002; Feng et al., 2017) and has been extended to multi-period settings (Hotz and Miller, 1993; Hotz et al., 1994).

5. In this paper we do not consider consumption so as to stay focused on the randomization utility flow.

6. See Bergman (1985) for economic and financial implications of Uzawa preferences.

7. See Dai et al. (2023) for a related result in the context of entropy-regularized reinforcement learning with log utility.

8. Fudenberg et al. (2015) discuss also APU as a way to tackle payoff uncertainty; see Section 5 therein and the related references cited. However, no model parameter unavailability is involved.

9. A companion paper (Dai et al., 2025a) studies the Merton problem from the RL and computational perspective. It introduces an auxiliary problem with a class of Gaussian policies solvable by RL algorithms and proves that its optimal policy can be used to recover the optimal (deterministic) policy to the original Merton problem. Additionally, several other works take this exploratory framework to devise RL algorithms applied to portfolio selection or stock execution problems, such as Wang and Zhou (2020); Wang et al. (2023); Huang et al. (2025). The purpose of this strand of literature is fundamentally different from that of the present paper.

10. We assume there is only one stock and one factor for notational simplicity. There is no essential difficulty in extending to the multi-stock/factor case.

11. Here we implicitly assume $\boldsymbol{\pi}$ to be absolutely continuous with respect to the Lebesgue measure so that the entropy can be properly defined.

12. One should, however, note that biased exploratory/randomized policies are common in the RL literature for various reasons. For example, the ϵ-greedy policy is a convex combination of the (classically) optimal one and a purely random policy; hence it is biased.

References

Agranov M, Ortoleva P (2017) Stochastic choice and preferences for randomization. *Journal of Political Economy* 125(1):40–68.

Agranov M, Ortoleva P (2025) Ranges of randomization. *Review of Economics and Statistics* 1–12.

Anderson SP, De Palma A, Thisse JF (1992) *Discrete Choice Theory of Product Differentiation* (MIT press).

Barberis N (2012) A model of casino gambling. *Management Science* 58(1):35–51.

Bender C, Thuan NT (2026) Continuous time reinforcement learning: a random measure approach. *Stochastic Processes and their Applications, forthcoming* .

Bergman YZ (1985) Time preference and capital asset pricing models. *Journal of Financial Economics* 14(1):145–159.

Berry ST (1994) Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* 242–262.

Ceballos-Lira MJ, Macias-Diaz JE, Villa J (2010) A generalization of Osgood's test and a comparison criterion for integral equations with noise. *arXiv preprint arXiv:1012.1843* .

Chacko G, Viceira LM (2005) Dynamic consumption and portfolio choice with stochastic volatility in incomplete markets. *The Review of Financial Studies* 18(4):1369–1402.

Dai M, Dong Y, Jia Y (2023) Learning equilibrium mean-variance strategy. *Mathematical Finance* 33(4):1166–1212.

Dai M, Dong Y, Jia Y, Zhou XY (2025a) Data-driven Merton's strategies via policy randomization. *arXiv preprint arXiv:2312.11797* .

Dai M, Jin H, Kou S, Xu Y (2021) A dynamic mean-variance analysis for log returns. *Management Science* 67(2):1093–1108.

Dai M, Sun Y, Xu ZQ, Zhou XY (2025b) Learning to optimally stop diffusion processes, with financial applications. *Management Science, forthcoming* .

Dianetti J, Dumitrescu R, Ferrari G, Xu R (2025) Entropy regularization in mean-field games of optimal stopping. *arXiv preprint arXiv:2509.18821* .

Duffie D, Epstein LG (1992) Stochastic differential utility. *Econometrica* 353–394.

Dwenger N, Kübler D, Weizsäcker G (2018) Flipping a coin: Evidence from university applications. *Journal of Public Economics* 167:240–250.

El Karoui N, Peng S, Quenez MC (1997) Backward stochastic differential equations in finance. *Mathematical Finance* 7(1):1–71.

Epstein LG, Zin SE (1989) Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57(4):937–969.

Feng G, Li X, Wang Z (2017) On the relation between several discrete choice models. *Operations Research* 65(6):1516–1525.

Fudenberg D, Iijima R, Strzalecki T (2015) Stochastic choice and revealed perturbed utility. *Econometrica* 83(6):2371–2409.

Guo X, Xu R, Zariphopoulou T (2022) Entropy regularization for mean field games with learning. *Mathematics of Operations Research* 47(4):3239–3260.

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

Han X, Wang R, Zhou XY (2023) Choquet regularization for continuous-time reinforcement learning. *SIAM Journal on Control and Optimization* 61(5):2777–2801.

He XD, Hu S, Obłój J, Zhou XY (2017) Path-dependent and randomized strategies in barberis' casino gambling model. *Operations Research* 65(1):97–103.

Hey JD, Carbone E (1995) Stochastic choice with deterministic preferences: An experimental investigation. *Economics Letters* 47(2):161–167.

Hotz VJ, Miller RA (1993) Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies* 60(3):497–529.

Hotz VJ, Miller RA, Sanders S, Smith J (1994) A simulation estimator for dynamic models of discrete choice. *The Review of Economic Studies* 61(2):265–289.

Hu S, Obłój J, Zhou XY (2023) A casino gambling model under cumulative prospect theory: analysis and algorithm. *Management Science* 69(4):2474–2496.

Hu Y, Imkeller P, Müller M, et al. (2005) Utility maximization in incomplete markets. *The Annals of Applied Probability* 15(3):1691–1712.

Huang Y, Jia Y, Zhou XY (2025) Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. *arXiv preprint arXiv:2412.16175* .

Jia Y (2024) Continuous-time risk-sensitive reinforcement learning via quadratic variation penalty. *arXiv preprint arXiv:2404.12598* .

Jia Y, Ouyang D, Zhang Y (2026) Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *SIAM Journal on Control and Optimization, forthcoming* .

Kydland FE, Prescott EC (1982) Time to build and aggregate fluctuations. *Econometrica* 1345–1370.

Liu J (2007) Portfolio selection in stochastic environments. *The Review of Financial Studies* 20(1):1–39.

Luce RD, et al. (1959) *Individual Choice Behavior*, volume 4 (Wiley New York).

Machina MJ (1985) Stochastic choice functions generated from deterministic preferences over lotteries. *The Economic Journal* 95(379):575–594.

Mattsson LG, Weibull JW (2002) Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior* 41(1):61–78.

McFadden D (1974) The measurement of urban travel demand. *Journal of Public Economics* 3(4):303–328.

McFadden D (2001) Economic choices. *American Economic Review* 91(3):351–378.

Merton RC (1969) Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics* 247–257.

Miao B, Zhong S (2018) Probabilistic social preference: how Machina's mom randomizes her choice. *Economic Theory* 65(1):1–24.

Permana Y (2020) Why do people prefer randomisation? an experimental investigation. *Theory and Decision* 88(1):73–96.

Tang W, Zhang YP, Zhou XY (2022) Exploratory HJB equations and their convergence. *SIAM Journal on Control and Optimization* 60(6):3191–3216.

Uzawa H (1968) Time preference, the consumption function, and optimum asset holdings. *Value, capital and growth: Papers in honor of Sir John Hicks*, 485–504 (The University of Edinburgh Press, Edinburgh).

Wachter JA (2002) Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets. *Journal of Financial and Quantitative Analysis* 37(1):63–91.

Wang B, Gao X, Li L (2023) Reinforcement learning for continuous-time optimal execution: actor-critic algorithm and error analysis. *Available at SSRN* .

Wang H, Zariphopoulou T, Zhou XY (2020) Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research* 21(198):1–34.

Wang H, Zhou XY (2020) Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical Finance* 30(4):1273–1308.

Zhao R, Sun X, Tresp V (2019) Maximum entropy-regularized multi-goal reinforcement learning. *International Conference on Machine Learning*, 7553–7562 (PMLR).

Ziebart BD, Maas AL, Bagnell JA, Dey AK (2008) Maximum entropy inverse reinforcement learning. *AAAI*, volume 8, 1433–1438 (Chicago, IL, USA).

Appendix A: Different Temperature Schemes

In this section we discuss two alternative temperature schemes for weighing the entropy utility, and explain the drawbacks of these formulations compared with our recursive formulation.

A.1. Constant Temperature

With the exploratory dynamics given in (5), if we were to use a constant temperature as in Wang et al. (2020) with the objective function

$$\mathbb{E}\left[\int_t^T \lambda\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + U(W_T^{\boldsymbol{\pi}})\Big|W_t^{\boldsymbol{\pi}} = w, X_t = x\right],$$

then the associated HJB equation would be

$$\begin{aligned}&\frac{\partial V}{\partial t} + \sup_{\boldsymbol{\pi}}\Bigg\{\Big(r + \big(\mu(t,x) - r\big)\operatorname{Mean}(\boldsymbol{\pi})\Big)wV_w + \frac{1}{2}\sigma^2(t,x)\Big(\operatorname{Mean}(\boldsymbol{\pi})^2 + \operatorname{Var}(\boldsymbol{\pi})\Big)w^2V_{ww}\\&+ m(t,x)V_x + \frac{1}{2}\nu^2(t,x)V_{xx} + \rho\nu(t,x)\sigma(t,x)\operatorname{Mean}(\boldsymbol{\pi})wV_{wx} + \lambda\mathcal{H}(\boldsymbol{\pi})\Bigg\} = 0,\end{aligned} \tag{27}$$

with the terminal condition $V(T,w,x) = U(w) = \frac{w^{1-\gamma}-1}{1-\gamma}$.

Similar to Wang et al. (2020), we can solve the maximization problem on the left-hand side of (27) and apply the verification theorem to conclude that the optimal policy is a normal distribution with mean

$$\frac{\big(\mu(t,x)-r\big)V_w}{-\sigma^2(t,x)wV_{ww}} + \frac{\rho\nu(t,x)V_{wx}}{-\sigma(t,x)wV_{ww}},$$

and variance

$$\frac{\lambda}{-\sigma^2(t,x)w^2V_{ww}}.$$

Plugging the above into (27), the equation becomes

$$\begin{aligned}&\frac{\partial V}{\partial t} + rwV_w + m(t,x)V_x + \frac{1}{2}\nu^2(t,x)V_{xx} + \frac{\lambda}{2}\log 2\pi\\&- \frac{\Big(\big(\mu(t,x)-r\big)V_w + \rho\nu(t,x)\sigma(t,x)V_{wx}\Big)^2}{2\sigma^2(t,x)V_{ww}} + \frac{\lambda}{2}\log\frac{\lambda}{-\sigma^2(t,x)w^2V_{ww}} = 0.\end{aligned} \tag{28}$$

To our best knowledge, this PDE generally admits neither a separable nor a closed-form solution due to a lack of the homothetic property. As a result, analytical forms of the optimal value function and optimal policy are both unavailable, making it hard to carry out further theoretical analysis such as a comparative study and sensitivity analysis.

A.2. Wealth-Dependent Temperature

In order to obtain a simpler HJB equation, we could consider a wealth-dependent temperature parameter. For example, for CRRA utility, we could take $\lambda(t,w,x) = \lambda w^{1-\gamma}$ where $\gamma \neq 1$. The derivations of (27) and (28) are similar by replacing λ with $\lambda w^{1-\gamma}$. With this wealth-dependent weighting scheme, the problem becomes homothetic in wealth with degree $1-\gamma$; hence the value function admits the form $V(t,w,x) = \frac{w^{1-\gamma}v(t,x)}{1-\gamma} + g(t)$, where g and v satisfy

$$(1-\gamma)g' - (1-\gamma)g = \frac{\partial v}{\partial t} + mv_x + \frac{\nu^2}{2}v_{xx},\ g(T,x) = -\frac{1}{1-\gamma},$$
$$\frac{\partial v}{\partial t} + (1-\gamma)rv + mv_x + \frac{\nu^2}{2}v_{xx} + \lambda\frac{1-\gamma}{2}\log\frac{2\pi}{\gamma\sigma^2 v} + \frac{(1-\gamma)\big((\mu-r)v + \rho\nu\sigma v_x\big)^2}{2\gamma\sigma^2 v},\ v(T,x) = 1.$$

Consequently, the optimal policy is a normal distribution with mean $\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)v_x(t,x)}{\gamma\sigma(t,x)v(t,x)}$ and variance $\frac{\lambda}{\gamma\sigma^2(t,x)v(t,x)}$.

The major difference between this formulation and the recursive one is that there is an extra term $v(t,x)$ in the denominator of the optimal variance of the former, which may result in an arbitrarily large randomization variance within a finite time period and consequently the non-existence of an optimal policy. For example, in the Black-Scholes case $(m,\nu \equiv 0)$, v is independent of x and satisfies an ODE whose solution will reach 0 with some choice of the coefficients. Specifically, $v(t) = \varphi(T-t)$, where φ satisfies an ODE:

$$\varphi'(\tau) = (1-\gamma)\Big((r + \frac{(\mu-r)^2}{2\gamma\sigma^2})\varphi(\tau) + \frac{\lambda}{2}\log\frac{2\pi\lambda}{\gamma\sigma^2} - \frac{\lambda}{2}\log\varphi(\tau)\Big),\ \varphi(0) = 1. \tag{29}$$

We have the following theorem.

THEOREM 3. *If $\gamma > 1$ and $r + \frac{(\mu-r)^2}{2\gamma\sigma^2} + \frac{\lambda}{2}\log\frac{2\pi\lambda}{\gamma\sigma^2} < 0$, then the solution to (29) reaches 0 in a finite time.*

Proof. Consider an ODE $y' = F(y)$ with the initial condition $y(0) = 1$, where $F(x) = ax + b - c\log x$ with some constants a, b and c. Let $y(\tau) = e^{-z(\tau)}$. Then z satisfies

$$z' = -e^z F(e^{-z}) = -a - be^z - cze^z,\ z(0) = 0. \tag{30}$$

If $F(x) < 0$ for any $x \in (0,1]$ or z is increasing, then applying the Osgood test (c.f., Ceballos-Lira et al. 2010) we conclude that the solution of (30) explodes at time $\tau_e = -\int_0^\infty \frac{1}{a+be^z+cze^z}\mathrm{d}z$. The desired result now follows from the fact that $y \to 0$ is equivalent to $z \to \infty$. Q.E.D.

Appendix B: CARA Utility

The RPU also works for the CARA utility, for which we use the dollar amount invested in the risky asset as the control (portfolio) variable, and denote by $\boldsymbol{\pi}$ the corresponding probability-density-valued control. The wealth process under $\boldsymbol{\pi}$ is

$$\mathrm{d}W_t^{\boldsymbol{\pi}} = [rW_t^{\boldsymbol{\pi}} + (\mu(t,X_t)-r)\,\mathrm{Mean}(\boldsymbol{\pi}_t)]\,\mathrm{d}t + \sigma(t,X_t)\left[\mathrm{Mean}(\boldsymbol{\pi}_t)\mathrm{d}B_t + \sqrt{\mathrm{Var}(\boldsymbol{\pi}_t)}\mathrm{d}\bar{B}_t\right]. \tag{31}$$

Consider the following regularized objective function:

$$J_t^{\boldsymbol{\pi}} := \mathbb{E}\left[\int_t^T -\lambda J_s^{\boldsymbol{\pi}}\mathcal{H}(\boldsymbol{\pi}_s)\mathrm{d}s + U(W_T^{\boldsymbol{\pi}})\Big|\mathcal{F}_t\right], \tag{32}$$

where $U(x) = -\frac{1}{\gamma}e^{-\gamma x}$. Under this recursive weighting scheme, the optimal policy is given by

$$\boldsymbol{\pi}^*(t,x) = \mathcal{N}\left(\Big(\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} - \frac{\rho\nu(t,x)}{\sigma(t,x)}u_x(t,x)\Big)e^{-r(T-t)}, \frac{\lambda}{\gamma^2\sigma^2(t,x)}e^{-2r(T-t)}\right),$$

where u satisfies

$$\begin{aligned} &u_t + m(t,x)u_x + \frac{1}{2}\nu^2(t,x)(-\gamma u_x^2 + u_{xx}) + \frac{\gamma\sigma^2(t,x)}{2}\left(\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} - \frac{\rho\nu(t,x)}{\sigma(t,x)}u_x\right)^2 \\ &+ \frac{\lambda}{2\gamma}\log\frac{2\pi\lambda}{\gamma^2\sigma^2(t,x)e^{2r(T-t)}} = 0, \quad u(T,x) = 0, \end{aligned}$$

and the associated optimal value function is $V(t,w,x) = -\frac{1}{\gamma}\exp(-\gamma u(t,x) - \gamma e^{r(T-t)}w)$. Hence we can develop a theory parallel to the CRRA utility. Details are left to the interested readers.

Appendix C: A BSDE Perspective

The optimal portfolio processes with parameter λ in the RPU problem can also be characterized by a BSDE.

THEOREM 4. *Suppose the following quadratic BSDE admits a unique solution*

$$\begin{cases} \mathrm{d}Y_t^{*(\lambda)} = -\left\{(1-\gamma)\Big[\frac{\lambda}{2}\log\frac{2\pi\lambda}{\gamma\sigma_t^2} + r + \frac{(\mu_t - r + \rho\sigma_t Z_t^{*(\lambda)})^2}{2\gamma\sigma_t^2}\Big] + \frac{1}{2}{Z_t^{*(\lambda)}}^2\right\}\mathrm{d}t + Z_t^{*(\lambda)}\mathrm{d}B_t^X, \\ Y_T^{*(\lambda)} = 0. \end{cases} \tag{33}$$

Then the optimal randomized control is given by

$$\boldsymbol{\pi}_t^{*(\lambda)} = \mathcal{N}\Big(\frac{\mu_t - r + \rho\sigma_t Z_t^{*(\lambda)}}{\gamma\sigma_t^2}, \frac{\lambda}{\gamma\sigma_t^2}\Big).$$

The solution to BSDE (33) corresponds to the PDE (15) in Theorem 1. The process $Y^{*(\lambda)}$ can be interpreted as an auxiliary process in the martingale duality theory, which stipulates that

$$\left(V_t^{\boldsymbol{\pi}} + \int_0^t \lambda\mathcal{H}(\boldsymbol{\pi}_s)\big[(1-\gamma)V_s^{\boldsymbol{\pi}} + 1\big]\mathrm{d}s\right)e^{Y_t^{*(\lambda)}}$$

is a supermartingale for any portfolio control $\boldsymbol{\pi}_t$ and a martingale when $\boldsymbol{\pi}_t = \boldsymbol{\pi}_t^{*(\lambda)}$, where $V_t^{\boldsymbol{\pi}} = \frac{{w_t^{\boldsymbol{\pi}}}^{1-\gamma}}{1-\gamma} e^{Y_t^{*(\lambda)}} - \frac{1}{1-\gamma}$ is the value function evaluated along the optimal wealth trajectories. Moreover, when $\lambda = 0$, the above result reduces to the classical one which was first derived by Hu et al. (2005).

The BSDE (33) offers important insights about the connection between the classical problem ($\lambda = 0$) and the RPU ($\lambda > 0$), some of them consistent with the unbiasedness discussion in Subsection 3.2. First, the only difference between the two problems is the extra term, $\frac{(1-\gamma)\lambda}{2} \log \frac{2\pi\lambda}{\gamma\sigma_t^2}$, in the driver of (33). This term becomes a deterministic function of t when $\gamma = 1$ (log-utility) or when σ_t is deterministic, in which case $Z^{*(\lambda)} = Z^{*(0)}$. This is because, in general, if (Y_t, Z_t) satisfies a BSDE $\mathrm{d}Y_t = -f(t, X_t, Z_t)\mathrm{d}t + Z_t \mathrm{d}B_t^X$ with $Y_T = F(X_T)$, then $(Y_t + C(t), Z_t)$ solves the BSDE $\mathrm{d}\tilde{Y}_t = -\big[c(t) + f(t, X_t, \tilde{Z}_t)\big]\mathrm{d}t + \tilde{Z}_t \mathrm{d}B_t^X$ with $\tilde{Y}_T = F(X_T)$, where c is a given deterministic function of t and $C(t) = \int_t^T c(s)\mathrm{d}s$. Hence $Z = \tilde{Z}$ follows from the uniqueness of solution.

Therefore, the optimal solution of the RPU problem has a mean that coincides with the classical solution, or the former is unbiased. Incidentally, this is consistent with an earlier result of Dai et al. (2023) in the mean-variance analysis for the log utility. Second, when $\rho = 0$, i.e., the market factors evolve independently from the stock price, the optimal randomized policy is also unbiased, even if $Z^{*(\lambda)} \neq Z^{*(0)}$ in general. This is intuitive because hedging is not necessary in this case. A special case is when the market factor X is deterministic, in which case the source of randomness B_t^X vanishes leading to $Z^{*(\lambda)} = Z^{*(0)} = 0$ and hence the optimal randomization is unbiased. Otherwise, optimal policies are in general biased in stochastic volatility models when σ_t is a function of the stochastic factor process X_t and the power of the utility $\gamma \neq 1$.