Data-Driven Merton's Strategies via Policy Randomization

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

Yuchao Dong
School of Mathematical Sciences
Tongji University, Shanghai, China, Shanghai 200092, ycdong@tongji.edu.cn

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

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

We study Merton's expected utility maximization problem in an incomplete market, characterized by a factor process in addition to the stock price process, where all the model primitives are unknown. The agent under consideration is a price taker who has access only to the stock and factor value processes and the instantaneous volatility. We propose an auxiliary problem in which the agent can invoke policy randomization according to a specific class of Gaussian distributions, and prove that the mean of its optimal Gaussian policy solves the original Merton problem. With randomized policies, we are in the realm of continuous-time reinforcement learning (RL) recently developed in Wang et al. (2020) and Jia and Zhou (2022a,b, 2023), enabling us to solve the auxiliary problem in a data-driven way without having to estimate the model primitives. Specifically, we establish a policy improvement theorem based on which we design both online and offline actor–critic RL algorithms for learning Merton's strategies. A key insight from this study is that RL in general and policy randomization in particular are useful beyond the purpose for exploration – they can be employed as a technical tool to solve a problem that cannot be otherwise solved by mere deterministic policies. At last, we carry out both simulation and empirical studies in a stochastic volatility environment to demonstrate the decisive outperformance of the devised RL algorithms in comparison to the conventional model-based, plug-in method.

Key words: Merton's problem; incomplete market; randomized policy; reinforcement learning; policy evaluation; policy improvement; actor–critic learning

1. Introduction

Merton's expected utility maximization model (Merton, 1969) and its subsequent rich variants are central to continuous-time finance. The traditional paradigm for applying the Merton models

to practice follows the so-called "separation principle" (separation between estimation and optimization), also known as the "plug-in" method. Starting from a basic stock price model – be it the simplest Black–Scholes, a stochastic volatility model, or a jump-diffusion – an econometrician estimates the model parameters/primitives from historical data using statistical or machine learning methods and then passes on to an (optimization) theorist. The latter plugs in the estimated values to the resulting stochastic control problem and solves it (rarely) analytically or (commonly) numerically via solving Pontryagin's maximum principle conditions or Hamilton–Jacobi–Bellman (HJB) partial differential equations (PDEs). The endeavors of the econometrician and the theorist are thus *separated*: The former deals with estimation only, and the latter takes the estimated model as given and focuses on optimization. *Had* an infinite amount of data been available, this division of labor might work – suitable statistical/econometric methods ensure that reasonable correctness of the model can be validated and the primitives be estimated to the highest accuracy possible. However, in the context of financial markets and asset returns, it has been well documented that accurate estimates of certain parameters – predominantly the expected return – require an amount of data far beyond the history of financial markets (Merton, 1980; Luenberger, 1998). Even worse, a market is most likely non-stationary, defeating the stationarity assumption usually required by those econometric methods. Furthermore, Merton's strategies, if computable, are typically very sensitive to model primitives. As a result, estimation errors may amply propagate to the theorist's final solutions, rendering them irrelevant to practice.

By contrast, the modern reinforcement learning (RL) paradigm takes a *conceptually* and fundamentally different approach.[1] It still begins with a basic structural model underlying the data-generating process (e.g., a Markov chain or a diffusion process), but it does not assume the model parameters to be given and known, *nor does it attempt to estimate them*. Instead, RL tries to learn optimal policies or strategies *directly*, via first parameterizing a policy and then updating (learning) its parameters iteratively to improve the policy until optimality or near-optimality is achieved. The approach accomplishes this typically in three steps: 1) strategically exploring the unknown environment (e.g., a market) by trial-and-error: randomly experimenting different choices according to some carefully designed probability distribution (called a randomized or *stochastic* policy) and observing the responses (called reward or reinforcement signals) from the environment; 2) learning the value function of that stochastic policy based on the reward signals; and 3) improving the stochastic policy based on the learned value function. These steps are called respectively *exploration, policy evaluation*, and *policy improvement*, and the resulting algorithms are referred to as the *actor* (policy)–*critic* (value function) type in RL. So RL is end-to-end, model-free and

[1] RL has been predominantly studied for discrete-time Markov decision processes (MDPs); see, e.g., Sutton and Barto (2011) for a systematic account.

data-driven: It maps data to decision policies, skipping the middle step of estimating model primitives.[2]

The primary goal of randomized/stochastic policies is for exploration: randomization broadens search space and enables observation of counterfactual outcomes of alternative choices (that otherwise would have never been tried by non-randomized, deterministic policies) to better understand the interactions between actions and environment. However, this goal seems to be irrelevant to the Merton problem when the investor is a *small* investor (i.e., a price taker). As stock prices are exogenous to the small investor, he can compute the return of any portfolio without actually purchasing it (e.g., using a "paper portfolio"). Therefore, there appear to be no informational benefits to adopting stochastic policies or the RL approach for such a small investor.

This paper aims to argue otherwise and show how RL can still be used to solve Merton's problem efficiently and effectively in a model-free, data-driven way, *even* for small investors. Indeed, we construct an auxiliary problem that allows for a special class of stochastic policies – Gaussian policies to be specific with a particular variance function – to relax the original Merton problem. The problem is inspired by the stochastic relax control formulation first proposed by Wang et al. (2020) for continuous-time RL. We then prove that the mean of the optimal Gaussian policy to the auxiliary problem is the optimal policy to the original Merton problem. This in turn justifies and demonstrates the significance of this auxiliary problem and, by extension, the randomized approach. More importantly, once stochastic policies are engaged, we are then in the realm of RL and able to develop RL algorithms based on the general theory and algorithms established in Jia and Zhou (2022a,b, 2023) to solve the constructed auxiliary problem with Gaussian policies. In particular, we design an algorithm tailor made for the Merton problem with power utility functions by leveraging its homothetic properties to enhance efficiency. We prove the convergence of the proposed algorithm in the Black–Scholes market with the "optimal" convergence rate typical in the literature. Interestingly, we show that stochastic policies are indeed *necessary* for our algorithms to work because these algorithms do not update any deterministic policies. Intuitively, as stochastic policies degenerate into a point mass (a deterministic policy), the variance of the reinforcement signal becomes so large that it no longer guides any policy improvement.

Next, using a special stochastic volatility model considered in Liu (2007), we further demonstrate why RL is preferable over the plug-in method even if the model class is correctly specified (but whose coefficients are unknown). Other than the challenge in statistically estimating some of the

[2] Throughout this paper, by "model-free" we mean that we do not have access to the model primitives, although – as mentioned earlier – we do have a basic structural model such as a diffusion process as in this paper. By "data-driven" we mean that policies are learned by observable/computable data – both exogenous and endogenous – such as stock price and volatility processes as in this paper.

model primitives as discussed earlier, this model showcases another important yet subtle difficulty when applying the plug-in method: the forms of the model-based optimal policies may drastically depend on the constellations of the model primitives and some of them may be practically insensible or infeasible, yet a statistical method is typically unable to take those parameter constraints into consideration when estimating them. In other words, statistical methods focus on estimating the model without necessarily considering its implications on the subsequent decision-making step. As a result, model estimation errors can propagate to policy errors in a profound way. By contrast, the RL approach starts with a reasonable structure of policies and improves them within that class, thereby avoiding the issue.

Finally, we report and discuss the results of both simulation and empirical studies comparing the performances of our RL algorithms with those of the classical plug-in method and a naïve buy-and-hold strategy. We find that the RL methods exhibit a clear and consistent advantage in terms of robustness and all-round performance.

Related Literature

The original Merton problem (Merton, 1969) is under a Black–Scholes market setting. Subsequent studies involve more general and richer market models, e.g., ones in which instantaneous mean return and/or volatility are driven by additional random sources. The corresponding Merton problem has been studied in, to name but a few, Wachter (2002); Chacko and Viceira (2005); Liu (2007). The literature on the Merton problem has been primarily from the perspective of an economic agent who, having already had access to a correct market model, focuses on solving the portfolio selection problem and provides insights into how different market conditions affect optimal portfolio choices and asset prices. There are papers addressing the agent's incomplete information on the expected stock returns, and they either assume that the agent conducts Bayesian learning (e.g., Gennotte 1986; Pástor 2000; Cvitanić et al. 2006; Andrei and Hasler 2015), or take a robust control approach to consider the worst scenario among a model class (e.g., Hansen and Sargent 2001; Maenhout 2004; Hansen et al. 2006). However, the former Bayesian approach has been restricted to simple models to keep the Bayesian updating tractable for analysis and computation, and the latter robust approach crucially relies on specifying a class of models while the model uncertainty is not endogenously determined.[3] To our best knowledge, no paper systematically studies the Merton problem for an investor with minimum knowledge about a "model" who learns optimal choice in both offline and online settings. The present paper aims to fill this void – it tackles the problem in an incomplete market by developing interpretable and efficient algorithms that learn the optimal policy without knowing or trying to estimate the market specifications.

[3] Epstein and Schneider (2007) discuss how to incorporate learning into model ambiguity, but the analysis is not tractable for complex models if the Bayesian posterior is not explicitly available.

This paper relates to a strand of literature on machine learning for financial decision making, especially its applications to dynamic portfolio choice. Gao and Chan (2000) and Jin and El-Saawy (2016) formulate a Merton problem as a discrete-time MDP that allows only a finite number of decisions. They apply Q-learning algorithms with portfolio return as a reward with/without adjusting for risk. In contrast, we consider continuous-time and continuous state–action spaces to reflect more realistic trading patterns including high-frequency transactions and allocation of an arbitrary percentage of total wealth to risky assets.[4] There have been also attempts to employ deep neural networks to solve MDPs with continuous state–action spaces or stochastic control problems; see, e.g., Han and E (2016); Bachouch et al. (2021), and Duarte et al. (2024). However, these papers assume the models are completely known and apply neural networks only as a computational tool to solve the respective optimization problems. As such, their approaches are alternatives to the traditional simulation or PDE-based numerical methods, instead of providing end-to-end solutions that map data to decisions. On the other hand, there are works that directly learn *deterministic* trading policies via the so-called "empirical risk minimization (ERM)"; see, for example, Guijarro-Ordonez et al. (2021) for the one-period mean–variance model and Buehler et al. (2019) for dynamic hedging. However, ERM can only do offline learning as it inherently requires the data in the whole time horizon while our method permits both online and offline learning. Moreover, Reppen and Soner (2023) demonstrate that ERM tends to perform poorly with limited data sets and exhibits desired convergence only with sufficiently large data sets. In the present paper, we will also show (in Appendix B) that our RL algorithms perform better than ERM when the sample size is small.

In recent years, there has been an upsurge of interest in continuous-time RL with continuous state and action spaces, not only because many practical problems are continuous time by nature (e.g., autonomous driving, robot navigation, and ultra-high-frequency trading) but also because more analytical tools are available in the continuous setting for developing a rigorous theory. Works by Wang et al. (2020); Jia and Zhou (2022a,b, 2023) lay the theoretical foundation for the formulation and algorithm design for continuous-time RL. A central underpinning of this series of research is the *martingality*: the learning/updating of the parameters of both the actor and the critic is guided by maintaining the martingality for various stochastic processes. Applications of these general results include, to name just a few, Wang and Zhou (2020) for continuous-time mean–variance pre-committed portfolio choice, Dai et al. (2023) for mean–variance equilibrium policies, Wang et al. (2023) for liquidation and execution, and Guo et al. (2022) for mean-field games. However, they have been largely restricted to the class of linear–quadratic problems. The present paper is the first to apply continuous-time RL to utility-based portfolio selection.

[4] For ease of presentation, in this paper, we consider a market with only one risky asset (e.g., a market index fund), but our method can be readily generalized to multiple assets.

The rest of the paper is organized as follows. Section 2 introduces the model-free Merton problem as well as an auxiliary problem with stochastic policies. In Section 3 we conduct a theoretical analysis including the connection between the original problem and the auxiliary one, based on which we present both online and offline RL algorithms to learn the optimal policy. We also prove convergence of the offline algorithm for the special Black–Scholes setting. Next, we use a class of stochastic volatility models to illustrate the benefits of the RL methods in Section 4. A simulation study and an empirical analysis are presented in Section 5. Finally, Section 6 concludes. All the proofs and additional results/discussions are placed in the appendix.[5]

2. Problem Formulation

Throughout this paper, with a slight abuse of notation, we use either Z or Z_t to refer to a stochastic process $Z := \{Z_t\}_{t\in[0,T]}$, while Z_t may also refer to the value of the process at time t if it is clear from the context. We use $f(\cdot)$ or f to denote a function, and $f(x)$ to denote the value of the function f at x. For a function f with arguments (t, w, x), we use $\frac{\partial f}{\partial t}$, $f_w, f_x, f_{ww}, f_{wx}, f_{xx}$ to denote its first- and second-order partial derivatives with respect to the arguments. We use bold-faced $\boldsymbol{\pi}$ to denote various probability-density-function valued portfolio controls or policies, and $\pi \approx 3.14$ and $e \approx 2.72$ to denote the respective mathematical constants. For a probability density function $\boldsymbol{\pi}$ on $\mathbb{R}$, we denote its mean and variance by $\text{Mean}(\boldsymbol{\pi}) = \int_{\mathbb{R}} a\boldsymbol{\pi}(a)\mathrm{d}a$ and $\text{Var}(\boldsymbol{\pi}) = \int_{\mathbb{R}} a^2\boldsymbol{\pi}(a)\mathrm{d}a - \text{Mean}(\boldsymbol{\pi})^2$, respectively. Finally, we denote by $\mathcal{N}(a, b^2)$ the density function of a normal distribution with mean a and variance b^2, with $\mathcal{N}(a, 0)$ specializing to the Dirac mass at point a.

2.1. Market Environment and Investment Objective

There are two assets available for investment in a market: a risk-free asset (bond) with a constant interest rate r and a stock (or market index). The stock price process is observable, whose dynamic is governed by the following 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 is a scalar-valued Brownian motion, and the instantaneous return rate process $\mu_t \equiv \mu(t, X_t)$ and volatility process $\sigma_t \equiv \sigma(t, X_t)$ both depend on another observable stochastic market factor process X. We assume that X follows SDE:

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

where $\tilde{B}$ is another (scalar-valued) Brownian motion independent of B, and $\rho \in (-1, 1)$ is a constant that determines the correlation between the stock return and the change in the market factor. So

[5] The code to reproduce the numerical results in this paper is available at `https://www.dropbox.com/scl/fo/onrln1ggs3vl46aclgno9/AMFr2VV2U1mQRNSScVyi300?rlkey=t55jyru3y9u9xttzncpbsorpy&st=fj6csbtu&dl=0`

the market is in general incomplete. We only consider the Markovian model, i.e., $\mu(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$, $m(\cdot,\cdot)$, and $\nu(\cdot,\cdot)$ are deterministic and continuous functions of t and x such that equations (1)–(2) have a unique weak solution.

This market setup is similar to that of Dai et al. (2021), which covers many popular and incomplete market models as special cases, e.g., the Gaussian mean return model and the stochastic volatility model studied in Wachter (2002), Liu (2007), and Chacko and Viceira (2005) among others.

A (small) investor's actions are modeled as a scalar-valued adapted process $a = \{a_t\}_{t\in[0,T)}$, with a_t representing the fraction of total wealth invested in the stock at time t. The corresponding self-financing wealth process W^a then follows the SDE:

$$\frac{\mathrm{d}W_t^a}{W_t^a} = a_t\frac{\mathrm{d}S_t}{S_t} + (1-a_t)r\mathrm{d}t = [r + (\mu(t,X_t)-r)a_t]\mathrm{d}t + \sigma(t,X_t)a_t\mathrm{d}B_t, \quad W_0^a = w_0. \tag{3}$$

Note that the solvency constraint $W_t^a \geqslant 0$ a.s., for all $t \in [0,T]$, is satisfied automatically for any square integrable a. The Merton investment problem is to choose a to maximize the following expected utility of the terminal wealth:

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

where W_T^a is defined by (2)-(3) and $U(\cdot)$ is a utility function.

We focus on the constant relative risk aversion (CRRA) utility function in the main body of this paper, i.e., $U(w) = \frac{w^{1-\gamma}-1}{1-\gamma}$, where $1 \neq \gamma > 0$ is the relative risk aversion coefficient.[6]

2.2. Agent's Knowledge and Randomized Choices

The classical Merton problem is model-based, namely, all the model primitives are assumed to be known and given, and the problem is typically solved by dynamic programming and HJB equations, leading to a *deterministic* optimal (feedback) policy.

In this paper, however, we consider an agent who does not have knowledge about the market environment up to the diffusion structure presented in the previous subsection and is unable to form a proper prior on each model within the family specified in (1) and (2) or unable to do Bayesian update of beliefs on each model. This setting is motivated by the difficulty in computing Bayesian posterior on general functional spaces, as well as the difficulty in specifying priors.[7] The agent encounters multiple episodes of investment tasks, with the investment horizon T for each

[6] When $\gamma = 1$, the CRRA utility function becomes the logarithm function $U(w) = \log w$. A problem with log utility can be regarded as a special case of the mean–variance problem for log returns in Dai et al. (2023) and Jiang et al. (2022). Hence we restrict our attention to the case of $\gamma \neq 1$ in this paper.

[7] Even if each model is indexed by a finite-dimensional vector, it is already challenging in posterior computation beyond the conjugate family. See more review and discussion in Green et al. (2015).

episode. Within each episode, the time is indexed by $t \in [0, T]$, and at time t, the past trajectories of the stock–factor value process and the wealth–portfolio process up to time t within the current episode can be observed. For simplicity, we assume the relative risk aversion coefficient γ is known to the agent. Finally, for the particular approach employed in this paper we need to assume that the volatility process $G_t = \sigma(t, X_t)^2$ is observable. This assumption is premised upon the well-documented results that the volatility may be approximated accurately by VIX, option data, or high-frequency observation of stock returns.[8]

So the agent's task is to solve the Merton problem in a data-driven way, where data include only stock–factor–volatility processes, the agent's own wealth process under any given portfolio, and the risk aversion parameter, without knowledge of the forms of market coefficients $\mu(\cdot,\cdot), \sigma(\cdot,\cdot), m(\cdot,\cdot), \nu(\cdot,\cdot)$. This knowledge/information structure more accurately reflects an actual investor's knowledge rather than a hypothetical omniscient agent. Moreover, in reality, when faced with an unknown environment, humans tend to do trial-and-error to test various strategies (i.e. engage randomized policies) and learn from experience.[9]

Both the knowledge structure and the employment of stochastic policies are prevalent in the general RL literature. A distinctive feature of RL compared with standard optimization or statistics is that "data" can be *endogenous* and, hence, also part of the solutions. It is generally acknowledged that a policy in RL has two objectives: to learn the environment *relevant* to the optimization objective and to improve performance. The former is the demand for exploration while the latter for exploitation. The essence of RL is to strike the best exploration–exploitation balance, which is usually achieved by randomizing decisions, i.e., extending the policy space to include stochastic or randomized policies (or mixed strategies in game theory). It is randomization that generates endogenous data for learning.

Following Wang et al. (2020), we now reformulate the Merton problem with stochastic policies. An investor chooses her time-t action (portfolio) by sampling from a probability distribution $\boldsymbol{\pi}_t$, where $\{\boldsymbol{\pi}_t\}_{t\in[0,T]} =: \boldsymbol{\pi}$ is a distribution-valued process called a stochastic or exploratory *control*. The resulting *exploratory dynamic* of the wealth process is described by

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

[8] For example, instantaneous variance can be calculated accurately based on the realized variance with high-frequency observations (Barndorff-Nielsen and Shephard, 2002; Hansen and Lunde, 2006). Alternatively, it is possible to use the derivative price on realized variance (Carr et al., 2005) as a proxy for the instantaneous variance, such as VIX for S&P 500 index.

[9] It is interesting to note that taking randomized decisions is often observed in behavioral experiments (Agranov and Ortoleva, 2017) and considered as an integral part of human behaviors (Mattsson and Weibull, 2002; Swait and Marley, 2013). Stochastic policies are popular in analyzing (dynamic) discrete choices (Hotz and Miller, 1993), and our setting here is a natural extension to accommodate continuum choices.

where $\bar{B}$ is another Brownian motion that is independent of both B and $\tilde{B}$, characterizing the additional noises introduced into the wealth process due to randomization. Intuitively, (5) is the limit of equations where actions are sampled from the randomized policy $\boldsymbol{\pi}$ at discrete times. The derivation of (5) is analogous to that in Dai et al. (2023) and an informal explanation is provided in Appendix A. A rigorous proof of how (5) describes the wealth process under the random portfolio choices is presented in Jia et al. (2025).

Finally, we reiterate the important point about the need and interpretation of randomized decisions in the particular Merton problem *with a small investor* that differs from the general RL. The rationale of using randomization for exploration is to learn how the (unknown) environment reacts to a greater number of different decisions. This rationale is only valid when such a reaction is unknown a priori. For example, one will not observe the return of a slot machine (the counterfactual) unless actually playing it. However, in the setting of this paper with a small investor, how the environment (market) reacts to the agent's decision (portfolio choice) can be *deduced*, as shown in the first equation in (3). Hence, observing the counterfactual returns of alternative portfolios is possible without having to actually execute those portfolios to gain information about the market. Therefore, the primary motivation for engaging stochastic policies in this paper, as explained earlier, is technical more than informational. That said, randomization will become essential also for the latter reason when we are to extend our study to involve a large investor whose actions will affect the market and hence exploration–exploitation tradeoff becomes relevant.

2.3. An Auxiliary Problem with Gaussian Policies

Our purpose is to develop an approach to solve the classical Merton's problem (4) subject to (2) and (3) by bypassing the conventional statistical estimation methods. To this end, we propose an auxiliary problem that incorporates stochastic policies, and show that the solution to the original problem can be derived and computed through that of the auxiliary problem.

We first introduce the following class of Gaussian (feedback) policies indexed by $\lambda \geqslant 0$ with a specific form of variance:

Definition 1. A measurable, distribution-valued function $\boldsymbol{\pi}^{(\lambda)} : [0,T] \times \mathbb{R}_+ \times \mathbb{R} \to \mathcal{P}(\mathbb{R})$, where $\lambda \geqslant 0$, is called an admissible policy, if

(i) $\boldsymbol{\pi}^{(\lambda)}(\cdot|t,w,x) = \mathcal{N}\left(\boldsymbol{u}(t,w,x), \frac{\lambda}{\gamma\sigma(t,x)^2}\right)$ for some measurable function $\boldsymbol{u} : [0,T] \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$, where by convention $\mathcal{N}(\boldsymbol{u}(t,w,x),0)$ is the Dirac mass at $\boldsymbol{u}(t,w,x)$;

(ii) under $\boldsymbol{\pi}^{(\lambda)}$, (5) has a unique weak solution $\{W_t^{\boldsymbol{\pi}^{(\lambda)}}\}_{t\in[0,T]}$ satisfying $\mathbb{E}\left[\sup_{0\leqslant t\leqslant T} |U(W_t^{\boldsymbol{\pi}^{(\lambda)}})|\right] < \infty$.

Moreover, for a given $\lambda \geqslant 0$, denote the collections of all admissible policies by $\boldsymbol{\Pi}^{(\lambda)}$.

This class of Gaussian policies are inspired by certain entropy-regularized optimization problems; see Ziebart et al. (2008) for a discrete-time setting and Wang et al. (2020), Wang and Zhou (2020), and Dai et al. (2023) for continuous-time counterparts. The variance of such a policy is inversely proportional to the volatility of the stock prices and the agent risk aversion level. The exogenous parameter $\lambda \geqslant 0$ controls the additional randomness (arising from policy randomization) introduced to the system.

For any given $\lambda \geqslant 0$, the objective of our auxiliary problem is to maximize

$$J^{(\boldsymbol{\pi}^{(\lambda)})}(t,w,x) = \mathbb{E}\Big[U(W_T^{\boldsymbol{\pi}^{(\lambda)}}) \,|\, W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\Big], \tag{6}$$

with the *optimal* value function

$$V^{(\lambda)}(t,w,x) = \max_{\boldsymbol{\pi}^{(\lambda)} \in \boldsymbol{\Pi}^{(\lambda)}} \mathbb{E}\Big[U(W_T^{\boldsymbol{\pi}^{(\lambda)}}) \,|\, W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\Big], \ (t,w,x) \in [0,T] \times \mathbb{R}_+ \times \mathbb{R}. \tag{7}$$

Note that this auxiliary problem is different from the entropy-regularized problems studied in Wang and Zhou (2020), where the entropy of the policy is explicitly included in the objective functional.

3. Theoretical Analysis

3.1. Ground Truth Solution to the Auxiliary Problem

We first answer the question on the relation between the auxiliary problem (7) and the original one (4). It is straightforward, as in Wang et al. (2020), to derive that the optimal value function $V^{(\lambda)}$ satisfies the following HJB equation via dynamic programming for (7):

$$\begin{aligned} &\frac{\partial V^{(\lambda)}}{\partial t} + \sup_{\boldsymbol{u} \in \mathbb{R}} \Bigg\{ \Big(r + \big(\mu(t,x) - r\big)\boldsymbol{u}\Big) w V_w^{(\lambda)} + \frac{1}{2}\sigma^2(t,x)\Big(\boldsymbol{u}^2 + \frac{\lambda}{\gamma\sigma(t,x)^2}\Big) w^2 V_{ww}^{(\lambda)} \\ &+ m(t,x) V_x^{(\lambda)} + \frac{1}{2}\nu^2(t,x) V_{xx}^{(\lambda)} + \rho\nu(t,x)\sigma(t,x)\boldsymbol{u} w V_{wx}^{(\lambda)} \Bigg\} = 0, \end{aligned} \tag{8}$$

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

At first glance, equation (8) is a highly nonlinear PDE and appears hard to analyze. However, we can reduce it to a simpler PDE based on which the optimal stochastic policy can be explicitly represented.

THEOREM 1. *Suppose φ is a classical solution of the following PDE*

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + (1-\gamma)r + m(t,x)\varphi_x + \frac{1}{2}\nu^2(t,x)(\varphi_{xx} + \varphi_x^2) \\ &\quad + \frac{1-\gamma}{2\gamma}\left[\frac{\big(\mu(t,x)-r\big)^2}{\sigma^2(t,x)} + \frac{2\rho\big(\mu(t,x)-r\big)\nu(t,x)}{\sigma(t,x)}\varphi_x + \rho^2\nu^2(t,x)\varphi_x^2\right] = 0, \end{aligned} \tag{9}$$

with the terminal condition $\varphi(T,x)=0$*, and* φ *satisfies the regularity condition that* $\{e^{(1+\epsilon)\varphi(t,X_t)}\}_{t\in[0,T]}$ *is uniformly integrable for some* $\epsilon>0$*. Then*

$$V^{(\lambda)}(t,w,x)=\frac{w^{1-\gamma}\exp\{\varphi(t,x)-\lambda(1-\gamma)(T-t)/2\}-1}{1-\gamma} \tag{10}$$

is a classical solution to the HJB equation (8)*. Moreover,*

$$\boldsymbol{\pi}^{(\lambda)*}(t,x)=\mathcal{N}\left(\boldsymbol{u}^*(t,x),\frac{\lambda}{\gamma\sigma^2(t,x)}\right),\ \textit{with}\ \boldsymbol{u}^*(t,x)=\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)}+\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}\varphi_x(t,x), \tag{11}$$

is the optimal policy to the auxiliary problem (7) *subject to* (2) *and* (5)*. Furthermore,* $\boldsymbol{u}^*$ *is the optimal policy for the original Merton's problem* (4) *subject to* (2) *and* (3)*.*

Theorem 1 characterizes the optimal ground truth solution (i.e. the theoretical solution *assuming* all the model coefficients are known) to the auxiliary problem (7) via the PDE (9) and reveals that its mean is none other than the optimal solution of the original problem (4). Note that this a *theoretical* result not to be used to compute the solutions to either problems. Rather, its importance lies in its implication: *one can solve* (4) *via solving* (7). It in turn justifies our approach of employing a special class of Gaussian policies to recover the optimal solution of the original Merton problem. Moreover, Theorem 1 indicates that we can limit the admissible policies to only bivariate functions $\boldsymbol{u}$ of (t,x), thus greatly reducing the complexity in solving the auxiliary problem.

As we discussed earlier there is no informational motive to study the auxiliary problem with stochastic policies due to the small investor in question (while there is such a motive in the case of a large investor whose actions impact the asset prices). Engaging the auxiliary problem (7) is a *technical* approach to learn the optimal solution to the original problem, as stipulated by Theorem 1. What is more, we will show subsequently that (7) can be solved by a policy improvement algorithm, which does *not* work directly on (4).

However, taking randomized policies is not free, because the utility value decreases due to the additional randomness borne by a risk-averse agent. We now study this "cost" by comparing them to deterministic policies (i.e. those with $\lambda=0$) in terms of the *equivalent relative wealth loss* (ERWL) defined as follows.

Definition 2. We define the equivalent relative wealth loss $\mathrm{ERWL}(\boldsymbol{\pi}^{(\lambda)})$ of an admissible policy $\boldsymbol{\pi}^{(\lambda)}$ as $\Delta=\Delta(t,x)$ satisfying

$$J^{(\boldsymbol{\pi}^{(\lambda)})}(0,w,x)=V^{(0)}(0,w(1-\Delta),x).$$

So ERWL Δ is a percentage in wealth with which investor is indifferent between obtaining the ground truth value of the optimal deterministic policy with initial endowment $w(1-\Delta)$ and getting the value of the optimal randomized policy with initial endowment w. In other words, Δ is the relative cost the investor is willing to pay to engage stochastic policies.

COROLLARY 1. *The equivalent relative wealth loss of the λ-optimal stochastic policy $\boldsymbol{\pi}^{(\lambda)*}$ is a constant that only depends on λ and the length of the episode T. Specifically,* $\mathrm{ERWL}(\boldsymbol{\pi}^{(\lambda)*}) = 1-\exp\{-\lambda T/2\} \approx \lambda T/2 + O(\lambda^2 T^2)$.

Corollary 1 quantifies the loss of efficiency in the (relative) monetary term due to a randomized policy: A longer investment horizon or a larger λ incurs larger losses, which is clearly intuitive.

3.2. Reinforcement Learning Methods for Solving the Auxiliary Problem

With the class of Gaussian policies, we are in the realm of RL and thus able to apply/develop RL methods to solve the auxiliary problem. The basic idea follows the actor–critic approach developed for general stochastic control problems in Jia and Zhou (2022b), with a major modification for Merton's problem.

An actor–critic type algorithms learns the value function and the policy function alternatingly and iteratively. The critic refers to the policy evaluation stage (estimating the value function under the current policy), and the actor corresponds to the policy improvement stage (updating the policy guided by the value function). Theorem 1 informs that it suffices to learn two bivariate functions of (t,x), $\boldsymbol{u}^*$ for the policy and φ for the value function. These bivariate functions can be approximated by, e.g., a certain parametric form, linear spans of basis functions like polynomials, or neural networks. We will specify them later in our numerical study.

Now that our task reduces to learning the two functions $\boldsymbol{u}^*$ and φ that only depend on the time and the market factors, it turns out these functions possess nice properties that they are "closed" in the iterative procedures of policy evaluation–policy improvement, precisely stipulated by Theorem 2 below.

THEOREM 2. *(i) The value function under an admissible policy $\boldsymbol{\pi}^{(\lambda)}(\cdot|t,x) = \mathcal{N}\left(\boldsymbol{u}(t,x), \frac{\lambda}{\gamma\sigma(t,x)^2}\right)$ can be represented as*

$$J^{(\boldsymbol{\pi}^{(\lambda)})}(t,w,x) = \frac{w^{1-\gamma}\exp\{\bar{\varphi}(t,x) - \lambda(1-\gamma)(T-t)/2\} - 1}{1-\gamma}, \tag{12}$$

where $\bar{\varphi}$ satisfies the PDE

$$\begin{aligned} &\frac{\partial\bar{\varphi}}{\partial t} + (1-\gamma)r + m(t,x)\bar{\varphi}_x + \frac{1}{2}\nu^2(t,x)\left(\bar{\varphi}_{xx} + (\bar{\varphi}_x)^2\right) \\ &\quad + (1-\gamma)\left[(\mu(t,x)-r)\boldsymbol{u}(t,x) - \frac{\gamma}{2}\sigma(t,x)^2\boldsymbol{u}(t,x)^2 + \rho\sigma(t,x)\nu(t,x)\boldsymbol{u}(t,x)\bar{\varphi}_x\right] = 0, \end{aligned} \tag{13}$$

with terminal condition $\bar{\varphi}(T,x) = 0$.

(ii) Define a new policy

$$\tilde{\boldsymbol{\pi}}^{(\lambda)}(\cdot|t,x) = \mathcal{N}\left(\tilde{\boldsymbol{u}}(t,x), \frac{\lambda}{\gamma\sigma(t,x)^2}\right),\ \tilde{\boldsymbol{u}}(t,x) = \frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)} + \frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}\bar{\varphi}_x(t,x). \tag{14}$$

Then this new policy $\tilde{\boldsymbol{\pi}}^{(\lambda)}$ improves $\boldsymbol{\pi}^{(\lambda)}$: $J^{(\tilde{\boldsymbol{\pi}}^{(\lambda)})}(t,w,x) \geqslant J^{(\boldsymbol{\pi}^{(\lambda)})}(t,w,x)$ for all (t,w,x).

Theorem 2-(i) confirms that it is indeed sufficient to consider the specific form (10) for the critic because *any* value function is of that form. Theorem 2-(ii) entails a *policy improvement* theorem, which specifies a provably better policy over any given policy. We emphasize that this theorem is a *theoretical* result and cannot in itself be used to compute the optimal policy and value function as we do not have access to the coefficients of the PDE (13).

The next theorem, however, forms the foundation for the algorithms we are going to devise, by characterizing the value function associated with any given admissible policy as well as the improved policy that are both theoretically identified by Theorem 2.

THEOREM 3. *(i) Let $\lambda>0$, an admissible Gaussian policy $\boldsymbol{\pi}^{(\lambda)}$ and a continuous function $\hat{V}$ be given satisfying $\hat{V}(T,w,x)=U(w)$ and for every $(t_0,w_0,x_0)\in[0,T)\times\mathbb{R}_+\times\mathbb{R}$,*

$$\mathbb{E}\left[\int_{t_0}^{T}\xi(t,W_t^{a^{\boldsymbol{\pi}^{(\lambda)}}},X_t)\mathrm{d}\hat{V}(t,W_t^{a^{\boldsymbol{\pi}^{(\lambda)}}},X_t)\right]=0 \text{ for any measurable function } \xi,$$

where $(W^{a^{\boldsymbol{\pi}^{(\lambda)}}},X)$ is the wealth–factor process under a control $a^{\boldsymbol{\pi}^{(\lambda)}}$ sampled from $\boldsymbol{\pi}^{(\lambda)}$ with the initial $W_{t_0}^{a^{\boldsymbol{\pi}^{(\lambda)}}}=w_0$, $X_{t_0}=x_0$. Then $\hat{V}\equiv J^{(\boldsymbol{\pi}^{(\lambda)})}$, which is given in Theorem 2-(i).

(ii) Let $\lambda>0$, a continuous function $\hat{\boldsymbol{u}}$ and its associated policy $\hat{\boldsymbol{\pi}}^{(\lambda)}(\cdot|t,x)=\mathcal{N}\left(\hat{\boldsymbol{u}}(t,x),\frac{\lambda}{\gamma\sigma(t,x)^2}\right)$ be given satisfying for every $(t_0,w_0,x_0)\in[0,T)\times\mathbb{R}_+\times\mathbb{R}$,

$$\mathbb{E}\left[\int_{t_0}^{T}\eta(t,W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}},X_t)\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\mathrm{d}J^{(\boldsymbol{\pi}^{(\lambda)})}(t,W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}},X_t)\right]=0 \text{ for any measurable function } \eta,$$

where $(W^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}},X)$ is the wealth–factor process under a control $a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}$ sampled from $\hat{\boldsymbol{\pi}}^{(\lambda)}$ with the initial $W_{t_0}^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}}=w_0$, $X_{t_0}=x_0$. Then $\hat{\boldsymbol{u}}\equiv\tilde{\boldsymbol{u}}$, which is constructed in Theorem 2-(ii).

The two equations in Theorem 3 are types of martingale orthogonality conditions studied extensively in (Jia and Zhou, 2022a, Section 4.2) that lead to model-free, data-driven stochastic approximation algorithms to compute the value and policy functions by choosing appropriate classes of the "test functions" ξ and η; see the next subsection for details. Notably, Theorem 3-(ii) explains why a stochastic policy *needs* to be considered in our approach. When only deterministic policies are adopted (i.e., $\lambda=0$), the orthogonality condition in Theorem 3-(ii) holds trivially for any $\hat{\boldsymbol{u}}$ because $a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}\equiv\hat{\boldsymbol{u}}(t,X_t)$, hence becomes useless.

3.2.1. Data-Driven RL Algorithms Based on Theorem 2, we only need to learn the functions $\bar{\varphi}$ and $\tilde{\boldsymbol{u}}$ in order to determine the value function of a given stochastic policy and its improved policy, respectively. Denote by $\hat{\varphi}^{\psi}$ and $\hat{\boldsymbol{u}}^{\theta}$ the respective approximated functions of $\bar{\varphi}$ and $\tilde{\boldsymbol{u}}$, where (ψ,θ) are finite-dimensional parameters to be learned. Then the corresponding approximated value function and (improved) policy are

$$\hat{V}^{\psi}(t,w,x)=\frac{w^{1-\gamma}\exp\{\hat{\varphi}^{\psi}(t,x)-\lambda(1-\gamma)(T-t)/2\}-1}{1-\gamma},\ \hat{\boldsymbol{\pi}}^{\theta}(\cdot|t,x)=\mathcal{N}\left(\hat{\boldsymbol{u}}^{\theta}(t,x),\frac{\lambda}{\gamma\sigma^2(t,x)}\right). \tag{15}$$

Applying Theorem 3 to the above functions yields

$$\begin{cases} \mathbb{E}\left[\int_0^T \xi_t \mathrm{d}\hat{V}^{\psi}(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t)\right] = 0, \\ \mathbb{E}\left[\int_0^T \eta_t \dfrac{a_t^{\hat{\boldsymbol{\pi}}^\theta} - \hat{\boldsymbol{u}}^\theta(t, X_t)}{\lambda/(\gamma G_t)} \mathrm{d}\hat{V}^{\psi}(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t)\right] = 0, \end{cases} \tag{16}$$

for suitably chosen "test processes" $\xi_t = \xi(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t)$ and $\eta_t = \eta(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t)$, where $a_t^{\hat{\boldsymbol{\pi}}^\theta}$ is the portfolio sampled from $\hat{\boldsymbol{\pi}}^\theta$ at time t and $W^{a^{\hat{\boldsymbol{\pi}}^\theta}}$ is the observed wealth process satisfying the wealth equation (3) under the resulting portfolio process, and $G_t = \sigma(t, X_t)^2$. [10]

With specified test functions ξ, η, (16) becomes a coupled system of algebraic equations in (ψ, θ), where the coefficients can be computed by *observable* data. The system of equations is also known as the *moment conditions* or *estimating equations* in the literature of generalized method of moment (Hansen and Singleton, 1982) in econometrics. However, we emphasize that the critical difference between econometrics and RL lies in that data are both exogenous and endogenous and a part of the solution with the latter, because samples of portfolios and wealth both depend on the policy $\hat{\boldsymbol{\pi}}^\theta$ that needs to be learned.

In the RL literature, typical choices of the test processes are $\xi_t = \frac{\partial}{\partial \psi}\hat{V}^{\psi}(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t)$ and $\eta_t = \frac{\partial}{\partial \theta}\hat{\boldsymbol{u}}^\theta(t, X_t)$, leading to the so-called "TD(0)" type of algorithms; see e.g., Sutton and Barto (2011). However, there is no formal theory on the "optimal" choice of these processes. For our problem, we propose the following

$$\xi_t = \frac{\frac{\partial}{\partial \psi}\hat{\varphi}^{\psi}(t, X_t)}{(1-\gamma)\hat{V}^{\psi}(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t) + 1}, \ \eta_t = \frac{\frac{\partial}{\partial \theta}\hat{\boldsymbol{u}}^\theta(t, X_t)}{(1-\gamma)\hat{V}^{\psi}(t, W_t^{a^{\hat{\boldsymbol{\pi}}^\theta}}, X_t) + 1}, \tag{17}$$

which effectively replace the TD error term $\mathrm{d}\hat{V}_t^{\psi}$ with an adjusted, "relative" TD error $\frac{\mathrm{d}\hat{V}_t^{\psi}}{(1-\gamma)\hat{V}_t^{\psi}+1}$ in a conventional TD(0) algorithm. The reason for this adjustment is due to the homothetic property of the CRRA utility function. In particular, the wealth processes are typically growing and non-stationary, which may cause instability in the learning process. The purpose of the denominator in (17) is to normalize the wealth effect.

3.2.2. An Example: The Black–Scholes Market To illustrate the general results derived so far, let us consider the classical Black–Scholes market where there are a risk-free asset and a risky one with constant model coefficients, and there is no market factor. Theorem 1 then specializes to a simple solution with $\boldsymbol{u}^* = \frac{\mu - r}{\gamma\sigma^2}$ and $V^{(\lambda)}(t, w) = \frac{w^{1-\gamma}\exp\{[r + \frac{(\mu-r)^2}{2\gamma\sigma^2}](1-\gamma)(T-t) - \lambda(1-\gamma)(T-t)/2\} - 1}{1-\gamma}$. Once again, agent's knowledge includes the values of γ, T and σ, whereas λ is a fixed temperature

[10] Here, we have assumed that continuous sampling of $a_t^{\hat{\boldsymbol{\pi}}^\theta}$ is possible. In actual implementation, the integrals in (16) will be replaced by summations and the sampled wealth–factor process will be evaluated with the forward Euler scheme that requires sampling $a_t^{\hat{\boldsymbol{\pi}}^\theta}$ only at discrete times, as illustrated in the next subsection.

parameter. In particular, the mean return μ is unknown. This setting is consistent with a consensus that the stock expected return is more difficult if not impossible to estimate accurately using statistical methods; see e.g. Luenberger (1998).

Inspired by the (theoretical) ground truth optimal solution, we approximate the value and policy functions with two scalar parameters ψ and θ:

$$\hat{V}^{\psi}(t,w)=\frac{w^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}-1}{1-\gamma},\ \hat{\boldsymbol{u}}^{\theta}(t)=\theta,\ \hat{\boldsymbol{\pi}}^{\theta}(\cdot|t)=\mathcal{N}\left(\theta,\frac{\lambda}{\gamma\sigma^2}\right).$$

With the proposed test processes (17), the optimality conditions (16) now read

$$\begin{cases}\mathbb{E}\left[\displaystyle\int_0^T\frac{T-t}{(W_t^{a^{\hat{\boldsymbol{\pi}}^{\theta}}})^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}}\mathrm{d}\frac{(W_t^{a^{\hat{\boldsymbol{\pi}}^{\theta}}})^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}}{1-\gamma}\right]=0\\ \mathbb{E}\left[\displaystyle\int_0^T\frac{\gamma\sigma^2(a_t^{\hat{\boldsymbol{\pi}}^{\theta}}-\theta)}{\lambda(W_t^{a^{\hat{\boldsymbol{\pi}}^{\theta}}})^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}}\mathrm{d}\frac{(W_t^{a^{\hat{\boldsymbol{\pi}}^{\theta}}})^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}}{1-\gamma}\right]=0\end{cases}.\tag{18}$$

An informal analysis: optimality conditions. To better understand the conditions (18), we first present an informal analysis by ignoring the time discretization issue (i.e. assuming it is possible to continuously draw samples from a stochastic policy and collect observations, and to compute the integrals involved exactly).

Applying Itô's lemma to the term $\mathrm{d}\frac{(W_t^{a^{\hat{\boldsymbol{\pi}}^{\theta}}})^{1-\gamma}\exp\{\psi(T-t)-\lambda(1-\gamma)(T-t)/2\}}{1-\gamma}$ and using the wealth equation (3), we deduce that (18) is equivalent to

$$\begin{cases}\mathbb{E}\left[\displaystyle\int_0^T(T-t)\left(-\frac{\psi}{1-\gamma}+\frac{\lambda}{2}+r+(\mu-r)a_t^{\hat{\boldsymbol{\pi}}^{\theta}}-\frac{\gamma}{2}\sigma^2(a_t^{\hat{\boldsymbol{\pi}}^{\theta}})^2\right)\mathrm{d}t+\int_0^T(T-t)\sigma a_t^{\hat{\boldsymbol{\pi}}^{\theta}}\mathrm{d}B_t\right]=0\\ \mathbb{E}\left[\displaystyle\int_0^T\frac{\gamma\sigma^2}{\lambda}(a_t^{\hat{\boldsymbol{\pi}}^{\theta}}-\theta)\left(-\frac{\psi}{1-\gamma}+\frac{\lambda}{2}+r+(\mu-r)a_t^{\hat{\boldsymbol{\pi}}^{\theta}}-\frac{\gamma}{2}\sigma^2(a_t^{\hat{\boldsymbol{\pi}}^{\theta}})^2\right)\mathrm{d}t+\int_0^T\frac{\gamma\sigma^2}{\lambda}(a_t^{\hat{\boldsymbol{\pi}}^{\theta}}-\theta)\sigma a_t^{\hat{\boldsymbol{\pi}}^{\theta}}\mathrm{d}B_t\right]=0\end{cases}.\tag{19}$$

Because $a_t^{\hat{\boldsymbol{\pi}}^{\theta}}\sim\mathcal{N}(\theta,\frac{\lambda}{\gamma\sigma^2})$, the expectations in (19) can be explicitly calculated, yielding the following system of equations:

$$\frac{T^2}{2}\left(-\frac{\psi}{1-\gamma}+r+(\mu-r)\theta-\frac{\gamma}{2}\sigma^2\theta^2\right)=0,\quad T\gamma\sigma^2\left(\frac{\mu-r}{\gamma\sigma^2}-\theta\right)=0.\tag{20}$$

The solutions to these equations coincide with the theoretical ground truth solutions, which in turn implies that the optimality conditions (19) or (18) indeed lead to the *correct* solutions. Meanwhile, in this special case, (20) shows that the second equation regarding policy optimization (in θ) is decoupled from the first equation on policy evaluation (in ψ and θ). Given that we are mainly interested in finding the optimal policy, we shall therefore focus on the second equation only.

So, if the second equation of (19) can be perfectly computed and implemented without the need of discretization then it will yield the correct optimal policy solution, θ, right away. However,

the above informal analysis begs a puzzling question: What is the impact of the randomization measured by the temperature parameter λ? The second equation in (20) appears to be independent of λ: Then why do we still need to randomize with $\lambda > 0$? To resolve this puzzle, we have to conduct a formal analysis from the sampling perspective.

A formal analysis: impacts of discretization and randomization. We now present a formal analysis by incorporating sampling errors in our procedure.

For numerical implementation, the term inside the expectation, say $e(\psi, \theta)$, in the second equation of (18) needs to be estimated by samples collected with suitable time discretization for the integration and discrete sampling of the stochastic policy. More precisely, for equally spaced time grids $0 = t_0 < t_1 < \cdots < t_K = T$ with grid size $\Delta t = \frac{T}{K}$ and a given set of value and policy function parameters (ψ, θ), we denote by $\widehat{e(\psi,\theta)}$ the estimate of $e(\psi, \theta)$, computed by

$$\begin{aligned}\widehat{e(\psi,\theta)} = &\sum_{k=0}^{K-1} \frac{\gamma\sigma^2(a_{t_k}-\theta)}{\lambda(W_{t_k})^{1-\gamma}\exp\{\psi(T-t_k)-\lambda(1-\gamma)(T-t_k)/2\}}\frac{1}{1-\gamma} \\ &\times\Big[(W_{t_{k+1}})^{1-\gamma}\exp\{\psi(T-t_{k+1})-\lambda(1-\gamma)(T-t_{k+1})/2\} \\ &-(W_{t_k})^{1-\gamma}\exp\{\psi(T-t_k)-\lambda(1-\gamma)(T-t_k)/2\}\Big] \\ = &\sum_{k=0}^{K-1}\frac{\gamma\sigma^2(a_{t_k}-\theta)}{\lambda(1-\gamma)}\left[\left(\frac{W_{t_{k+1}}}{W_{t_k}}\right)^{1-\gamma}\exp\left\{[-\psi+\lambda(1-\gamma)/2]\,\Delta t\right\}-1\right],\end{aligned} \tag{21}$$

where $a_{t_k}|W_{t_k} \sim \mathcal{N}(\theta, \frac{\lambda}{\gamma\sigma^2})$, and on (t_k, t_{k+1}), W satisfies the wealth equation (3) with a constant portfolio a_{t_k}, i.e.,

$$\frac{\mathrm{d}W_t}{W_t} = a_{t_k}\frac{\mathrm{d}S_t}{S_t} + (1-a_{t_k})r\mathrm{d}t.$$

The impact of the time discretization is characterized by the following proposition.

PROPOSITION 1. *Suppose $\lambda > 0$ and $\Delta t < \min\{T, \frac{1}{\theta^2}, \frac{1}{|\psi|}\frac{1}{4\lambda|\gamma-1|}\}$. Then there exists a constant C that depends only on $\mu, r, \sigma, \gamma, T$ such that*

$$\left|\mathbb{E}\left[\widehat{e(\psi,\theta)}\right] - T\gamma\sigma^2(\theta^*-\theta)\right| \leqslant C(1+|\theta^2|+|\psi|+\lambda)\Delta t,$$

where $\theta^* = \frac{\mu-r}{\gamma\sigma^2}$. Moreover,

$$\mathrm{Var}\left[\widehat{e(\psi,\theta)}\right] \leqslant C\left(1+\frac{\theta^2}{\lambda}\right) + C\left(\frac{1+\psi^2+\theta^4}{\lambda}+\lambda\right)\Delta t.$$

While Proposition 1 confirms our derivation of the theoretical equivalence between the second equations of (18) (or (19)) and (20), it also shows there is a bias when we numerically compute the former due to time discretization and it gives an upper bound of the bias in terms of the grid size Δt and the strength of the randomization λ along with other parameters. The bound is linear

in Δt, consistent with the typical rate in numerical methods for simulating SDEs and computing integrals. It is also linear in λ; so a smaller level of randomization reduces the bias. Thus, from the bias-reducing perspective, besides a finer grid size, a smaller level of policy randomization helps.

On the other hand, Proposition 1 gives an upper bound for the variance of the learning signal samples. This bound will not vanish even if Δt shrinks to 0, implying that it is impossible to accurately compute the desired quantity with just a few trajectories (i.e. with a small dataset) even when continuous sampling is possible. The leading term consists of a constant and a term of the order λ^{-1}; so when the dataset size is small an elevated level of policy randomization helps reduce the variance.

The above analysis shows that randomization is indeed relevant for learning the optimal policy parameter θ in actual implementation. We should not pick a too large or too small λ in order to balance bias and variance. For a fixed $\lambda > 0$, the next theorem gives an algorithm to compute θ based on the discretized version of the second equation of (18) along with its convergence rate and the error bound of the expected equivalent relative wealth loss.

THEOREM 4. *Fix $\lambda > 0$ and consider the following update rule for the policy parameter θ:*

$$\theta_{n+1} = \Pi_{[-c_{n+1}, c_{n+1}]} \left(\theta_n + \ell_n \widehat{e_n(\psi, \theta)} \right), \tag{22}$$

where $\widehat{e_n(\psi,\theta)}$ is a sample given by (21) with grid size Δt_n and actor–critic parameters (θ_n, ψ_n), and $\Pi_K(\cdot)$ is the projection mapping onto a closed, convex set K. Assume that for any given $\eta_1 > 0, \eta_2 \in (0,1)$, there exist $n_0 \in \mathbb{N}$ and $M > 0$ such that for all $n \geqslant n_0$, $\ell_n = \frac{(1+\eta_1)}{(n+\eta_1)\eta_2\eta_1} < \frac{1}{T\gamma\sigma^2}$, $\Delta t_n \leqslant T\ell_n$, $c_n = \sqrt{\log n}$, and $|\psi_n| \leqslant M$. Then there exist constants C_1, C_2, C_3 that depend only on $\mu, r, \sigma, \gamma, T, n_0, M$, such that

$$\mathbb{E}\left[(\theta_{n+1} - \theta^*)^2\right] \leqslant C_1 \ell_n \log n \sim O\left(n^{-1} \log n\right),$$

and

$$\mathbb{E}\left[\mathrm{ERWL}(\hat{\boldsymbol{u}}^{\theta_n})\right] \leqslant \left(\max\{\frac{C_3 \lambda^{-1}}{1-\eta_2}, C_2\} \right) \ell_n \log n \sim O\left(n^{-1} \log n\right).$$

So, under the policy iteration algorithm (22) along with the chosen hyperparamters specified in the theorem, the algorithm converges and the expected equivalent relative wealth loss of the resulting RL policy converges to 0. Moreover, the L^2-convergence rate of the former and the convergence rate of the latter are both $O(n^{-1}\log n)$, matching the typical optimal convergence rate (ignoring the logarithmic factor). In particular, we obtain an error bound of the expected loss in terms of the strength of the randomization, which is of the order λ^{-1}. This result reconciles with Proposition 1 in terms of the variance of the learning signal. We illustrate the convergence rate under different values of λ in Figure 1. In the log-log scale plot, all curves eventually exhibit the

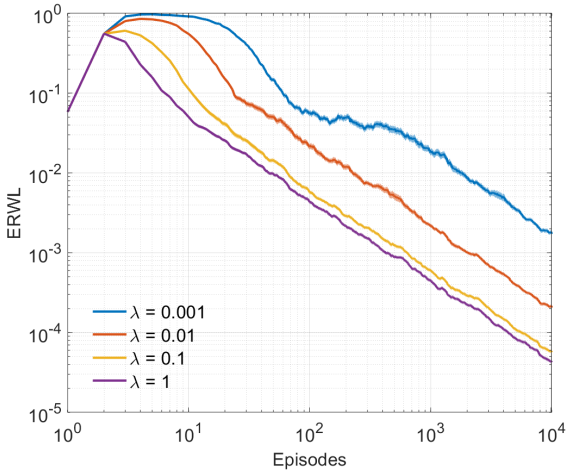

Figure 1 **The convergence of the learned policy (in terms of ERWL) under different temperature parameters λ. The horizontal (the number of episodes) and vertical (expected relative wealth loss) axes are both in log-scale. The shaded areas indicate the standard deviations of the estimated ERWLs. The results are based on 1000 times of independent simulation runs and 10,000 episodes of 1-year trajectory is used in each run. The model parameters are $\mu = 0.2, r = 0.02, \sigma = 0.3, \gamma = 3, T = 1$. The learning rate is $a_n = 10/(n+1)$ and the initial policy parameter is $\theta_0 = 0$. The projected region is taken as $c_n = \max\{10, \sqrt{\log(n+1)}\}$ and discretization size is $\Delta t_n = \min\{0.001, 10/(n+1)\}$.**

similar slope around -1, confirming the theoretical rate of convergence and the fact that error is reduced with a larger λ with small sample size.

Finally, note that our method does not permit $\lambda = 0$, and therefore always learns a stochastic policy. Alternative data-driven approaches, such as ERM, have been developed to directly learn a deterministic policy. We compare our method with ERM in a simulation study presented in Appendix B. The study demonstrates that the two have similar performance with abundant data (which requires a small $\lambda > 0$ for our method), but the latter is significantly worse when the sample size is small.

4. A Market with Stochastic Volatility

In this section, we present a stochastic volatility market environment, which is the setting for our subsequent numerical experiments, and discuss the advantages of RL over the classical plug-in approach.

A stochastic volatility model sets $\mu(t,x) = r + \delta x^{\frac{1+\alpha}{2\alpha}}$, $\sigma(t,x) = x^{\frac{1}{2\alpha}}$, $m(t,x) = \iota(\bar{x} - x)$, and $\nu(t,x) = \bar{\nu}\sqrt{x}$, where $\alpha \neq 0$. This is a fairly general model studied in Liu (2007) for the classical utility maximization problem and in Dai et al. (2021) for the classical equilibrium mean–variance problem for log returns. As before, we assume volatility process is observable/estimatable and γ is known, while the agent has no access to any other parameter values.

4.1. Classical Benchmark

For readers' convenience, we first review the relevant results of the classical model-based benchmark with $\lambda = 0$. The following lemma is taken from Liu (2007).

LEMMA 1. *Assuming the model primitives of the classical benchmark model satisfy*

$$\iota^2\gamma > (1-\gamma)(2\rho\iota\delta\bar{\nu} + \delta^2\bar{\nu}^2), \tag{23}$$

the optimal strategy is

$$\boldsymbol{u}^*(t,x) = \left(\frac{\delta}{\gamma} + \frac{\rho\bar{\nu}}{\gamma}A_1(t)\right)x^{\frac{\alpha-1}{2\alpha}} = \left(\frac{\delta}{\gamma} + \frac{\rho\bar{\nu}}{\gamma}A_1(t)\right)(\sigma^2(t,x))^{\frac{\alpha-1}{2}},$$

and the optimal value function is

$$V^{(0)}(t,w,x) = \frac{w^{1-\gamma}e^{A_1(t)x+A_0(t)} - 1}{1-\gamma},$$

where A_1, A_0 respectively satisfy the following ordinary differential equations (ODEs):

$$\begin{aligned} &A_1' - \iota A_1 + \frac{1}{2}\bar{\nu}^2A_1^2 + \frac{1-\gamma}{2\gamma}\big[\delta^2 + 2\rho\delta\bar{\nu}A_1 + \rho^2\bar{\nu}^2A_1^2\big] = 0,\ A_1(T) = 0,\\ &A_0' + (1-\gamma)r - \beta + \iota\bar{x}A_1 = 0,\ A_0(T) = 0. \end{aligned} \tag{24}$$

Indeed, the two ODEs in (24), under the condition (23), can be explicitly solved with the following solutions:

$$\begin{aligned} A_1(t) &= \frac{-\psi_1 + \psi_1 e^{\psi_0(T-t)}}{-\psi_2 + \psi_3 e^{\psi_0(T-t)}},\\ A_0(t) &= \psi_4(T-t) + \psi_5\log\frac{\big(-\psi_2 + \psi_3 e^{\psi_0(T-t)}\big)}{-\psi_2 + \psi_3}, \end{aligned} \tag{25}$$

where

$$\begin{aligned} \psi_0 &= -\frac{\sqrt{\iota^2\gamma - (1-\gamma)\delta\bar{\nu}(\delta\nu + 2\iota\rho)}}{\sqrt{\gamma}},\quad \psi_1 = \frac{(1-\gamma)\delta^2}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]},\\ \psi_{2,3} &= \frac{\iota\gamma - (1-\gamma)\delta\bar{\nu}\rho \pm \sqrt{\gamma}\sqrt{\iota^2\gamma - (1-\gamma)\delta\bar{\nu}(\delta\nu + 2\iota\rho)}}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]},\\ \psi_4 &= (1-\gamma)r + \iota\bar{x}\psi_3,\ \psi_5 = \frac{2\gamma\iota\bar{x}}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]}. \end{aligned}$$

The above analytical representations require specifications of the model parameters and, hence, cannot be used directly in our RL setting. However, they give specific functional *structures* of the value function and policies that are helpful for function approximations. We will employ them in our subsequent numerical experiments.

Lemma 1 shows that the optimal policy can be represented as a function of the stock volatility. Moreover, the elasticity of the instantaneous variance on the optimal policy is a constant, given by

$$\frac{\partial\boldsymbol{u}^*}{\partial\sigma^2}(t,x)\Big/\frac{\boldsymbol{u}^*(t,x)}{\sigma^2(t,x)} = \frac{\alpha-1}{2},$$

which represents the sensitivity of the portfolio with respect to the current (observed) volatility.

4.2. Pitfall of Model-Based Solution and Virtue of Reinforcement Learning

It is well known that solutions to Merton's problems are highly sensitive to model primitives (Merton, 1980), especially for the stochastic volatility model. The pitfall of the traditional model-based, first-estimate-then-optimize paradigm is twofold. First, the optimal solution depends on model primitives in a highly nonlinear way, as exemplified by (25), where ψ_i are complicated functions of the model parameters. This calls for an extremely accurate estimation of these functions, which may require unrealistically long historical data. Second, there is a technical assumption (23) in Lemma 1, which also appears in Kraft (2005). Such an assumption is to theoretically ensure the ODE system (24) to be well-posed. When the assumption is violated, the solutions of (24) have completely different forms:[11]

$$\begin{aligned} A_1(t) &= -\psi_0 + \psi_1 \tan\big(\psi_2(T-t) + \arctan\frac{\psi_0}{\psi_1}\big), \\ A_0(t) &= \psi_3(T-t) + \psi_4 \log\frac{\cos\big(\arctan\frac{\psi_0}{\psi_1} + \psi_2(T-t)\big)}{\cos\Big(\arctan\frac{\psi_0}{\psi_1}\Big)}, \end{aligned} \tag{26}$$

where

$$\begin{aligned} \psi_0 &= \frac{\iota\gamma - (1-\gamma)\delta\bar{\nu}\rho}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]}, \quad \psi_1 = \frac{\sqrt{\gamma}\sqrt{-\iota^2\gamma + (1-\gamma)\delta\bar{\nu}(\delta\nu + 2\iota\rho)}}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]}, \\ \psi_2 &= \frac{\sqrt{-\iota^2\gamma + (1-\gamma)\delta\bar{\nu}(\delta\nu + 2\iota\rho)}}{2\sqrt{\gamma}}, \quad \psi_3 = (1-\gamma)r - \beta - \iota\bar{x}\psi_0, \quad \psi_4 = -\frac{2\gamma\iota\bar{x}}{\bar{\nu}^2[\rho^2 + \gamma(1-\rho^2)]}. \end{aligned}$$

These functions will blow up to infinity periodically and thus do not lead to reasonable investment strategies. Yet, even if the true underlying market processes do satisfy (23), standard estimation procedures do not usually account for such a nonlinear and nonconvex constraint. As a consequence, the *estimated* model primitives may violate (23) so that the corresponding ODE system (24) may have solutions not in the same form as (25), and the resulting investment strategies may generate infinite leverage yielding infeasible numerical computations.

By contrast, RL bypasses model estimation and learns the optimal policy *directly*, thereby avoiding blow-up solutions described above arising from the traditional plug-in method. What RL learns or estimates is now the optimal policy itself rather than model primitives, based on performance rather than statistical properties. Specifically for the current stochastic volatility model, RL first determines the structures of the optimal policy and the value function through theoretical analysis, and then learns/updates the parameters in (25) through data and standard RL procedures such as policy evaluation and policy improvement.

[11] Because the ODE (24) is autonomous and separable, its general solution can be written as an indefinite integral: $T - t = \int_{A(t)}^{0} \frac{\mathrm{d}z}{\iota z - \frac{1}{2}\bar{\nu}^2 z^2 - \frac{1-\gamma}{2\gamma}[\delta^2 + 2\rho\delta\bar{\nu}z + \rho^2\bar{\nu}^2 z^2]}$. The form of the solution depends drastically on whether or not the quadratic algebraic equation $\iota z - \frac{1}{2}\bar{\nu}^2 z^2 - \frac{1-\gamma}{2\gamma}[\delta^2 + 2\rho\delta\bar{\nu}z + \rho^2\bar{\nu}^2 z^2] = 0$ has real roots.

We emphasize the importance of exploiting the special structure of a given problem for RL algorithm design. For instance, for the present problem, it follows from both the general result (Theorem 1) and the special one (Lemma 1) that we only need to consider the following class of Gaussian policies:

$$\mathrm{Var}(\boldsymbol{\pi}^*(t,x)) = \frac{\lambda}{\gamma\sigma^2(t,x)},\ \mathrm{Mean}(\boldsymbol{\pi}^*(t,x)) = \left[\frac{\delta}{\gamma} + \frac{\rho\bar{\nu}}{\gamma}A_1(t)\right](\sigma^2(t,x))^{\frac{\alpha-1}{2}}, \tag{27}$$

which we intentionally express in terms of the instantaneous variance σ^2. Note that the policy variance depends only on σ^2, whereas the mean depends on σ^2 as well as other model primitives through A, a function of time t only. Due to this special structure obtained through theoretical analysis, we can determine the policy variance without incurring any extra training or estimation so long as a proxy of σ^2 is available/observable. Naturally, we still need to learn the mean of the policy, but the learning will be amply simplified.

4.3. Numerical Procedure

By Itô's formula, the instantaneous variance process $G_t := \sigma^2(t, X_t) = X_t^{1/\alpha}$ satisfies

$$\mathrm{d}G_t = \left[(\frac{\iota\bar{x}}{\alpha} + \frac{(1-\alpha)\bar{\nu}^2}{2\alpha^2})G_t^{1-\alpha} - \frac{\iota}{\alpha}G_t\right]\mathrm{d}t + \frac{\bar{\nu}}{\alpha}G_t^{1-\alpha/2}[\rho\mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}\tilde{B}_t].$$

We now replace X with G as a state variable (the other state variable is wealth W) which is observable. In the current setting we do not need to assume the factor X to be observable.[12] On the one hand, to our best knowledge, we do not know of any statistical method tailored for estimating the coefficients of (1) and (2) using time series $\{(S_t, G_t)\}_{t\in[0,T)}$, except for the naïve MLE method that demands huge computational cost and large amount of data to reach desired accuracy. By contrast, RL methods take G_t as inputs to learn directly the function approximators for optimal policy without having to estimate the market model.

Consequently, in view of (15), we now consider the approximated value function and policy as

$$\hat{V}^\psi(t,w,g) = \frac{w^{1-\gamma}\exp\{\hat{\varphi}^\psi(t,g) - \lambda(1-\gamma)(T-t)/2\} - 1}{1-\gamma},\ \hat{\boldsymbol{\pi}}^\theta(\cdot|t,g) = \mathcal{N}\left(\hat{\boldsymbol{u}}^\theta(t,g), \frac{\lambda}{\gamma g}\right),$$

where the argument g stands for the observable instantaneous variance $G_t = \sigma(t,X_t)^2$.

There are two ways to further parameterize these actor–critic functions. Inspired by Lemma 1, especially the expressions in (25), we can parameterize the value function of a given policy as

$$\hat{V}^\psi(t,w,g) = \frac{w^{1-\gamma}}{1-\gamma}\exp\left(A_1^\psi(t)g^{\psi_6} + A_0^\psi(t) - \frac{\lambda(1-\gamma)(T-t)}{2}\right) - \frac{1}{1-\gamma},$$

[12] Note that knowing the instantaneous variance G_t is not equivalent to knowing the market factor X_t because α is unknown.

where

$$A_1^{\psi}(t) = \frac{-\psi_1 + \psi_1 e^{\psi_0(T-t)}}{\psi_2 + \psi_3 e^{\psi_0(T-t)}},\ A_0^{\psi}(t) = \psi_4(T-t) + \psi_5 \log \frac{\psi_2 + \psi_3 e^{\psi_0(T-t)}}{\psi_2 + \psi_3}, \tag{28}$$

with $\psi \in \mathbb{R}^7$ whose components are $\psi_0, \psi_1, \cdots, \psi_6$. Moreover, in view of both Theorem 1 and Lemma 1, we parameterize the policy by

$$\hat{\boldsymbol{\pi}}^{\theta}(a|t,g) = \frac{1}{\sqrt{\frac{2\pi\lambda}{\gamma g}}} \exp\left\{-\frac{\gamma g}{2\lambda}\left(a - g^{\theta_6}[\theta_4 + \theta_5 A_1^{\theta}(t)]\right)^2\right\},$$

where A^{θ} is parameterized by a set of *different* parameters $(\theta_0, \theta_1, \theta_2, \theta_3)$ but in the *same* form as A_1^{ψ} in (28). In total, $\theta \in \mathbb{R}^7$ consists of entries $\theta_0, \theta_1, \cdots, \theta_6$.

An alternative way is to engage neural networks. We can parameterize the value function by

$$V^{\psi}(t,w,g) = \frac{w^{1-\gamma}\exp\{(T-t)NN^{\psi}(t,g)\}}{1-\gamma} - \frac{1}{1-\gamma},$$

and the policy by

$$\boldsymbol{\pi}^{\theta}(a|t,g) = \frac{1}{\sqrt{\frac{2\pi\lambda}{\gamma g}}} \exp\left\{-\frac{\gamma g}{2\lambda}\left(a - NN^{\theta}(t,g)\right)^2\right\},$$

where NN^{ψ} and NN^{θ} are two neural networks with suitable dimensions of ψ and θ. Note these neural network constructions have also taken advantage of the theoretical results.

Finally, we use the stochastic approximation algorithm to search for the root to the estimating equations (16) with the test functions chosen as (17), where all the processes and integral are approximated via discretization in a way similar to that described in Subsection 3.2.2. We summarize these procedures as Algorithms 1 and 2 in both online and offline settings.

5. Numerical Studies

5.1. Simulation with Synthetic Data

A key advantage of a simulation study is that we have the ground truth ("omniscient") solutions available to compare against the learning results. In this subsection we report our numerical study with synthetic data, where sample paths of stock price and instantaneous variance process are simulated using the Euler–Maruyama scheme. The data are generated from the "3/2 model" with $\delta = \mu - r$ and $\alpha = -1$. In this case, the stock price and factor dynamics are

$$\frac{\mathrm{d}S_t}{S_t} = \mu \mathrm{d}t + \frac{1}{\sqrt{X_t}}\mathrm{d}B_t, \quad \mathrm{d}X_t = \iota(\bar{x} - X_t)\mathrm{d}t + \bar{\nu}\sqrt{X_t}(\rho \mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}\tilde{B}_t).$$

It is a typical non-affine stochastic volatility model proposed by Drimus (2012). In the classical case ($\lambda = 0$), the optimal policy is given by $\boldsymbol{u}^*(t,x) = (\mu - r)x/\gamma + \rho\bar{\nu}A(t)x/\gamma$.

The parameters are modified from the estimated values in Chacko and Viceira (2005), namely, $\delta = 0.2811$, $r = 0.02$, $\alpha = -1$, $\iota = 0.1374$, $\bar{x} = 35$, $\bar{\nu} = 0.9503$, and $\rho = 0.5241$.[13] The risk aversion

[13] Under the originally estimated parameters in Chacko and Viceira (2005), the buy-and-hold is almost the optimal policy. To avoid this coincidence, we modify some parameters so that different methods produce distinct results.

Algorithm 1 Online-Incremental Learning Algorithm

Inputs: initial wealth w_0, initial stock price s_0, initial instantaneous variance g_0, horizon T, time step Δt, number of mesh grids K, initial learning rates l_θ, l_ψ and learning rate schedule function $\ell(\cdot)$ (a function of the number of episodes), functional form of parameterized value function $\hat{V}^\psi(\cdot,\cdot,\cdot)$, functional form of parameterized policy function $\hat{\boldsymbol{\pi}}^\theta(a|t,g)$, interest rate r, risk aversion coefficient γ, temperature parameter λ.

Required program: market simulator $(s',g') = Market_{\Delta t}(t,s,g)$ that takes current time, stock price, and instantaneous variance, (t,s,g), as inputs and generates stock price s' and instantaneous variance g' at time $t+\Delta t$ as outputs.

Learning procedure:

Initialize θ, ψ.

for episode $j = 1$ **to** ∞ **do**

Initialize $k = 0$. Observe initial wealth w_0, initial stock price s_0, and initial instantaneous variance g_0. Store $w_{t_k} \leftarrow w_0$, $s_{t_k} \leftarrow s_0$, $g_{t_k} \leftarrow g_0$.

while $k < K$ **do**

Generate action $a_{t_k} \sim \boldsymbol{\pi}^\psi(\cdot|t_k, g_{t_k})$.

Apply a_{t_k} to market simulator $(s',g') = Market_{\Delta t}(t_k, s_{t_k}, g_{t_k})$, and observe new state s', g'.

Store $s_{t_{k+1}} \leftarrow s'$, $g_{t_{k+1}} \leftarrow g'$.

Compute current wealth $w_{t_{k+1}} = w_{t_k} + w_{t_k} a_{t_k} \frac{s_{t_{k+1}}}{s_{t_k}} + w_{t_k}(1-a_{t_k}) r \Delta t$.

Compute

$$\delta = \frac{\hat{V}^\psi(t_{k+1}, w_{t_{k+1}}, g_{t_{k+1}}) - \hat{V}^\psi(t_k, w_{t_k}, g_{t_k})}{(1-\gamma)\hat{V}^\psi(t_k, w_{t_k}, g_{t_k}) + 1}.$$

Update θ and ψ by

$$\psi \leftarrow \psi + \ell(j) l_\psi \delta \frac{\partial \hat{V}^\psi}{\partial \psi}(t_k, w_{t_k}, g_{t_k}).$$

$$\theta \leftarrow \theta + \ell(j) l_\theta \delta \frac{\partial}{\partial \theta} \log \hat{\boldsymbol{\pi}}^\theta(a_{t_k}|t_k, g_{t_k}).$$

Update $k \leftarrow k+1$

end while

end for

coefficient is taken as $\gamma = 3$, which is a common value estimated from the aggregated growth and consumption data (Kydland and Prescott, 1982). We further set the investment horizon $T = 1$ (year), the initial wealth $w_0 = 1$, initial market factor $x_0 = \bar{x}$, the temperature parameter $\lambda = 0.1$, and the time discretization step size $\Delta t = \frac{1}{250}$. To mimic a real-world scenario, we generate a training dataset with daily data for 20 years, and each time we randomly sample a consecutive

Algorithm 2 Offline Learning Algorithm

Inputs: initial wealth w_0, initial stock price s_0, initial instantaneous variance g_0, horizon T, time step Δt, number of mesh grids K, initial learning rates l_θ, l_ψ and learning rate schedule function $\ell(\cdot)$ (a function of the number of episodes), functional form of parameterized value function $\hat{V}^\psi(\cdot,\cdot,\cdot)$, functional form of parameterized policy function $\hat{\boldsymbol{\pi}}^\theta(a|t,g)$, interest rate r, risk aversion coefficient γ, temperature parameter λ.

Required program: market simulator $(s',g') = Market_{\Delta t}(t,s,g,a)$ that takes current time, stock price, and instantaneous variance (t,x,g) as inputs and generates stock price s' and instantaneous variance g' at time $t+\Delta t$ as outputs.

Learning procedure:

Initialize θ, ψ.

for episode $j = 1$ **to** ∞ **do**

Initialize $k = 0$. Observe initial wealth w_0, initial stock price s_0, initial instantaneous variance g_0. Store $w_{t_k} \leftarrow w_0$, $s_{t_k} \leftarrow s_0$, $g_{t_k} \leftarrow g_0$.

while $k < K$ **do**

Generate action $a_{t_k} \sim \boldsymbol{\pi}^\psi(\cdot|t_k, g_{t_k})$.

Apply a_{t_k} to market simulator $(s',g') = Market_{\Delta t}(t_k, s_{t_k}, g_{t_k})$, and observe new state s', g'.

Store $s_{t_{k+1}} \leftarrow s'$, $g_{t_{k+1}} \leftarrow g'$.

Compute current wealth $w_{t_{k+1}} = w_{t_k} + w_{t_k} a_{t_k} \frac{s_{t_{k+1}}}{s_{t_k}} + w_{t_k}(1-a_{t_k}) r \Delta t$.

Compute and store

$$\delta_{t_k} = \frac{\hat{V}^\psi(t_{k+1}, w_{t_{k+1}}, g_{t_{k+1}}) - \hat{V}^\psi(t_k, w_{t_k}, g_{t_k})}{(1-\gamma)\hat{V}^\psi(t_k, w_{t_k}, g_{t_k}) + 1}.$$

Update $k \leftarrow k+1$

end while

Update θ and ψ by

$$\psi \leftarrow \psi + \ell(j) l_\psi \sum_{k=0}^{K-1} \delta_{t_k} \frac{\partial \hat{V}^\psi}{\partial \psi}(t_k, w_{t_k}, g_{t_k}).$$

$$\theta \leftarrow \theta + \ell(j) l_\theta \sum_{k=0}^{K-1} \delta_{t_k} \frac{\partial}{\partial \theta} \log \hat{\boldsymbol{\pi}}^\theta(a_{t_k}|t_k, g_{t_k}).$$

end for

subsequence from that dataset with a length of 1 year as one episode for training (i.e. for updating the parameters (ψ,θ)). The batch size for training is kept the same as $m_{train} = 16$. The initial learning rate is set to be 0.01 and decays as $l(j) = j^{-1/2}$. In total we carry out 2000 episodes for learning. On the other hand, the test set contains $N_{test} = 10^4$ independent wealth trajectories, each

generated from an episode having one-year length under the (deterministic) mean policy with the parameter θ learned from the training. We reiterate that, in view of Theorem 1, we use stochastic policies for training and the mean of the learned stochastic policy for testing.

For the simulation study we apply throughout the offline algorithm, Algorithm 2, for learning/-training. Moreover, we implement two versions of function approximation for execution. One uses the specific parametric forms motivated by the theoretical solutions, denoted by "This Paper – Specific Form". The other one applies neural networks, denoted by "This Paper – Neural Network". In particular, for the latter we use two three-layer neural networks to approximate the value function and the stochastic policy, respectively. We then compare these algorithms with the ground truth ("Omniscient") as well as two other methods. The first one is a naïve buy-and-hold policy ("B-H") that only holds the risky asset throughout without rebalance. It can also be regarded as the benchmark for investment if the risky asset is a market index (e.g. S&P 500). The second is an estimate-and-plug-in policy based on the stochastic volatility model ("Est-SV") with the analytical solutions given by Lemma 1. We employ a maximum likelihood estimation approach to estimate the parameters of the 3/2 model using the training set (with the length of 20 years).[14]

We use two performance criteria to compare the different methods. The first one is the average utility value on the test set. Specifically, given a deterministic policy obtained from training under a given method, we apply it to the test set and obtain N_{test} independent one-year wealth trajectories. Denote the terminal wealth of these trajectories by $\hat{W}_T^{(i)}$, $i = 1, 2, \cdots, N_{test}$. Then the average utility is $\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{\hat{W}_T^{(i)^{1-\gamma}} - 1}{1-\gamma}$. The second criterion is the equivalent relative wealth loss, computed by finding Δ such that $V^{(0)}(0, w_0(1-\Delta), x_0) =$ average payoff on the test set, where $V^{(0)}$ is the optimal value function of the classical Merton problem under the true model.

Finally, to examine statistical significance of the proposed methods, we repeat the above simulation runs for 100 times with different random seeds. That is, for every simulation run, we first generate training data with a 20-year length and then apply each method to the same training data. After having obtained a learned/estimated policy through training, we calculate its two performance criteria on the same testing data, which consists of 10000 independent 1-year trajectories. The results, including both the averages and standard errors of these 100 simulation runs, are summarized in the upper panel of Table 1. The B-H policy is independent of any model or learning specifications, yielding about 2/3 of the omniscient utility value and 18.28% loss of initial endowment. On average, with the correctly specified model class, the Est-SV policy performs much better than B-H, generating 95.37% of the optimal utility and 2.96% loss in wealth. The RL algorithm

[14] The estimation is carried out by maximizing the likelihood function by the gradient ascent algorithm. The log-likelihood function of the data is approximated based on the Euler–Maruyama discretization of the SDEs, which coincides with the actual data generation process used in the simulation study.

with the specific parametric form outperforms B-H and Est-SV by considerable margins, with very small losses of utility (97.03% of the optimal utility) and relative wealth (2.16%). By contrast, the RL with neural networks performs worse, and on a par with Est-SV. We have done extensive experiments and observed that this finding is robust with respect to the structures of the neural networks used. The reason behind the discrepancy between the two RL methods is that in the simulation study, the specific parametric method uses the *correct* form of the optimal policy that corresponds to the true underlying data-generating process, while neural networks do not use much such structural information. In the experiment presented here, the size of the training dataset is relatively small; so approximation with the correct form performs better than using general neural networks, the latter likely over-fitted. Indeed, the training set contains only 20-year data so the distribution of the training set may considerably differ from the theoretical distribution due to sampling errors. Moreover, as we take 2000 episodes for training, the data in those episodes overlap and are hence not mutually independent. To this point, we provide extra numerical results in Appendix C based on a huge amount of data, where new and *independent* trajectories are generated in each training episode. In that experiment, neural networks perform equally well as the other RL method.

Next we examine the robustness of our algorithms with respect to the observable volatility process, motivated by the considerations that in practice one only has access to an approximated value of the volatility, and/or that the stochastic volatility model is wrongly specified. To this end, we construct a noisy observation $\tilde{G}_t = (\sqrt{G_t} + 0.02\xi_t)^2$, where $\xi_t \sim \mathcal{N}(0,1)$ are i.i.d. at (daily) observation times. This construction applies to both the training and testing datasets. This implies that the observed volatility signals deviate from the true one by 2% on average, and the agent only observes $(S_t, \tilde{G}_t)$. The corresponding comparisons across various methods are presented in the lower panel of Table 1. Compared with the previous results, the specific parametrization RL method still performs well and is only slightly worse than the case with exact volatility, while the neural network based method yields almost identical performance to its non-noisy counterpart. Note that B-H does not rely on volatility; so it has identical results as before. It is most noteworthy, however, that the performance of Est-SV drops dramatically owing to the contaminated data, which once again confirms the sensitivity (and drawbacks) of the conventional plug-in methods. By contrast, our RL methods are "model-parameter-free" and learn policies directly, resulting in a much more robust performance.

5.2. Empirical Study with Real Market Data

We study dynamic allocation between the S&P 500 index and a money market account with $r = 2\%$ risk-free interest rate to illustrate the performance of our RL algorithms in the real market.

Table 1 Performance comparison of different methods under 100 simulation runs. We compute the average utility value under each policy based on independent 10^4 one-year wealth trajectories, and then use the formula in Definition 2 to convert average utility to equivalent relative wealth loss. B-H stands for the buy-and-hold policy, and Est-SV for the estimate-and-plug-in policy. Each policy other than Omniscient and B-H is obtained from a simulated training set of daily data for 20 years, and the simulation is repeated with 100 independent runs. The numbers in the bracket indicate the standard errors.

Volatility	Method	Utility	Equivalent Relative Wealth Loss
	Omniscient	0.303	0
	B-H	0.201	18.28%
Exact	This Paper - Specific Form	0.294	2.16%
		(0.001)	(0.24%)
	This Paper - Neural Network	0.290	3.04%
		(0.001)	(0.21%)
	Est-SV	0.289	2.96%
		(0.002)	(0.35%)
Noisy	This Paper - Specific Form	0.286	3.84%
		(0.002)	(0.32%)
	This Paper - Neural Network	0.290	3.04%
		(0.001)	(0.21%)
	Est-SV	0.238	28.01%
		(0.005)	(0.21%)

S&P 500 is one of the most actively traded indices and its option market is also highly liquid.[15] Therefore, we can easily obtain volatility-related data from the market. In particular, VIX is an index administered by CBOE (Chicago Board Options Exchange) since 1990 based on option prices that reflects the market-priced average forward-looking volatility of the S&P 500 index, and is widely considered to be a proxy of the instantaneous volatility. VIX itself is a traded future with options written on it. In our empirical study, we take the S&P 500 index as the risky asset and VIX as a proxy for its volatility, both observable. We take data from 1990-01-01 to 2025-2-28 and use the first 10 years (up to 1999-12-31) as the pre-training period and leave the rest as the testing period. During the former period, we apply our offline algorithm, Algorithm 2, to learn the parameters (ψ, θ) and set the learned ones as the initial parameters for the latter period. Then we use the online algorithm, Algorithm 1, to learn and implement optimal Merton's strategies as we go. We fix our initial wealth on 2000-01-01 to be 1 dollar and take the risk aversion parameter as $\gamma = 3$. The benchmark policies to compare against are still the buy-and-hold (B-H) and the estimate-and-plug-in (Est-SV). We do not allow leverage or borrowing for all the policies under comparison; so if a method suggests taking leverage or short selling, then we truncate the portfolio value to be in the interval $[0, 1]$.

[15] There are many mutual funds and ETFs tracking S&P 500, including The SPDR S&P 500 ETF (SPY).

Also, to avoid seasonality that depends on the investment horizon, we consider only time-invariant policies, which can be viewed as the limit when the time-to-maturity approaches infinity. This seems reasonable given that we have a rather long testing period. Note that for a stochastic volatility model, such time-invariant policies still result in time-variant portfolios via the (time-variant) instantaneous volatility. The form of the optimal policy in (27) then becomes

$$\text{Mean}(\boldsymbol{\pi}^*(g)) = C_1 g^{C_2}, \quad \text{Var}(\boldsymbol{\pi}^*(g)) = \frac{\lambda}{\gamma g}$$

for some constants C_1, C_2.

The Est-SV is implemented as follows. First, we also restrict to time-invariant policies. We use a rolling window with a length of 10 years to estimate the model parameters and then plug-in to the analytical form of the optimal solution under the SV model. To save computational cost and avoid re-estimating the whole model every day, we only update the estimation of model parameters by maximizing the log-likelihood function along the gradient ascend direction for one step during the testing period.[16]

A comparison of different methods is summarized in Table 2, in terms of several commonly used metrics including (annualized) return, volatility, Sharpe ratio, (downside) semi-volatility, Sortino ratio, Calmar ratio, maximum drawdown, and recovery time. Among them Sharpe ratio is the most important and popular criterion because the essential goal of the Merton problem is to maximize the risk-adjusted return. We observe the two RL methods outperforms the other two methods in all the metrics except the annualized return (B-H has 5.6%, slightly over 5.3% by RL with neural networks). In particular, RL with neural networks beats the other methods by significant margins in most criteria including the Sharpe ratio. Moreover, the two RL methods have remarkably smaller maximum drawdowns during the whole period in which the market experienced a 56.8% drawdown. Even more notably, their recovery times are decisively and overwhelmingly shorter.[17] These observations indicate that RL strategies not only perform strongly but also robustly, and react to the environment change and make adjustment very quickly.

While Table 2 gives a glance of overall and average performance comparison over 25 years, we now inspect the wealth trajectories under different policies, presented in Figure 2. It is clear that RL with neural networks outperforms (in terms of portfolio worth) all the others prior to around 2020, taken over by B-H only after 2020. However, both RL portfolios are much less volatile than B-H, corroborating the findings of Table 2. In particular, the RL strategies considerably and consistently

[16] This is analogous to the online updating in RL algorithms. The construction of the log-likelihood function involved is described in Footnote 14.

[17] This significantly shorter recovery times of RL strategies have also been observed in the continuous-time mean–variance setting Huang et al. (2022, 2024).

Table 2 Comparison of out-of-sample performances of different methods from January 2000 to February 2025. We report the (annualized) return (Rtn), volatility (Vol), Sharpe ratio, (downside) semi-volatility (Semi-Vol), Sortino ratio, Calmar ratio, maximum drawdown (MDD), and recovery time (RT). The risk-free interest $r = 0.02$.

Method	Rtn	Vol	Sharpe	Semi-Vol	Sortino	Calmar	MDD	RT
This Paper: Specific Form	0.036	**0.075**	0.217	**0.057**	0.284	0.065	**0.251**	282
This paper: Neural Network	0.053	0.115	**0.289**	0.086	**0.385**	**0.098**	0.339	**202**
B-H	**0.056**	0.194	0.187	0.142	0.256	0.064	0.568	1376
Est-SV	0.025	0.099	0.055	0.074	0.073	0.012	0.440	4248

beat the other two during the first 10 years, 2000–2010. Recall that this is an extremely volatile period, including two bear markets, the dot com bubble burst in the early 2000s and the financial crisis during 2007–2008.[18]

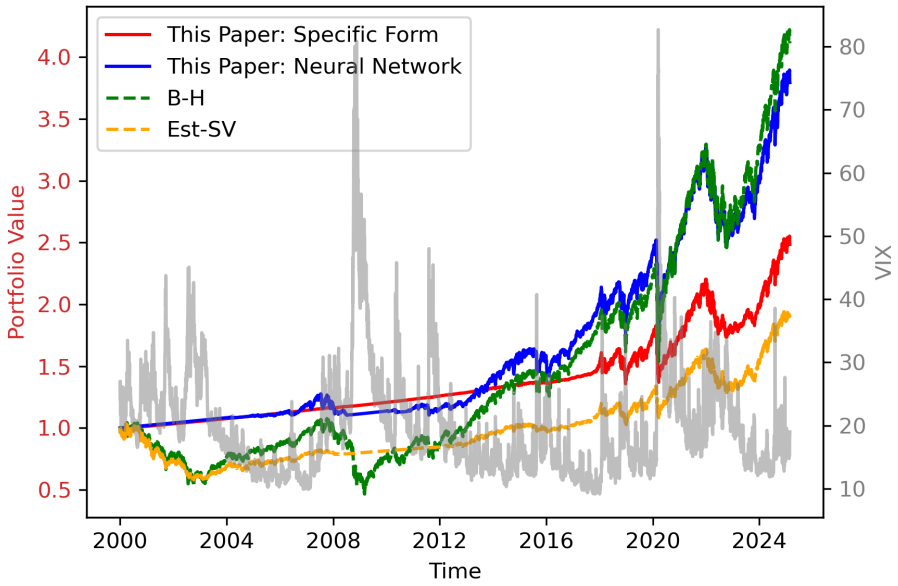

Figure 2 Wealth trajectories of portfolios under different policies. The gray plot is the VIX index whose vertical axis is on the right. The other plots are the trajectories of the portfolio values under different methods and are all normalized to 1 initially.

We further examine the proportions of wealth invested in the risky asset under different strategies, depicted in Figure 3. An interesting observation is that in the first half of the 2000–2010 overall bear period, the two RL-portfolios, especially the one with specific form, do not hold much risky asset as opposed to Est-SV. It demonstrates how the RL approach fundamentally differs from the traditional plug-in approach: Est-SV estimates model parameters statistically based on the

[18] The robustness, especially the outperformance during bear markets, of RL strategies devised from continuous-time theory is also documented for mean–variance portfolio choice; see Huang et al. (2022, 2024).

market data in the previous 10 years (1990–1999) that had a positive risk premium and that were characteristically different from those in the early 2000s. By contrast, RL learns portfolio strategies through real-time interactions with the market and pivots timely to more conservative ones after the market pivots. On the other hand, all the methods detect buying signals after 2010, while RL with neural networks is the first to react and start to gradually overweigh the risky asset.

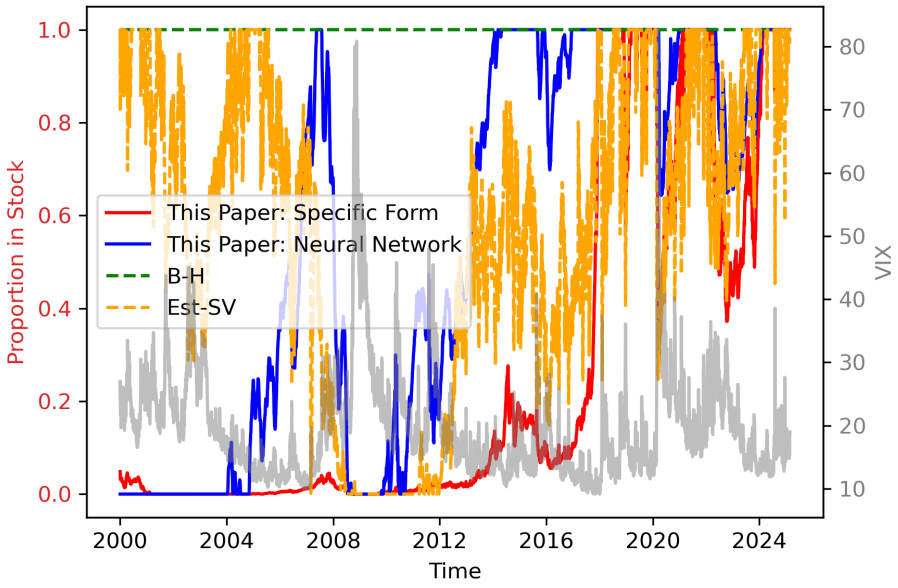

Figure 3 Trajectories of risky proportions under different policies. In our study, proportions invested in S&P 500 are restricted to be between 0 and 1. The gray curve is the VIX index whose vertical axis is on the right. The other curves are the trajectories of the proportions of the risky investment under different methods. The initial allocations of all the methods (except B-H) are based on the pre-training period from 1990 to 1999.

The empirical results indicate, decisively, that the RL methods are superior to the conventional Est-SV method in all fronts. As for the competition between the two RL algorithms, the one with neural networks outperforms the other by a good margin, contrary to the results from the simulation study. The reason is because the specific parametric form we adopt follows from a specific 3/2 model, which is almost certainly not valid in the real market, while the flexible structure of neural networks helps to identify other possible forms of investment strategies by exploiting as much as possible the VIX signals.

6. Conclusions

RL, as one of the cutting-edge technologies in artificial intelligence, has been applied to various fields. The central component of RL is exploration, which is carried out by policy randomization to broaden the action space aiming at understanding the interactions between an unknown environment and actions for improving and optimizing decision-making.

Applications of RL in finance however, especially in portfolio choice, is still in the early innings. One of the questions is that, unlike the bandits problem, stock data are exogenous and hence there is no need to "explore" – to actually try different portfolios to see the outcomes – if the market impact is ignored. In other words, there is no exploration–exploitation tradeoff because no extra information is gained by trial and error.

In this paper, we argue otherwise. To wit, we show that RL including policy randomization can go beyond the conceptual role of exploration; it can actually be used also as a *technical* tool to solve a "model-free" problem that otherwise cannot be solved satisfactorily by the conventional model-based methods. We do this in the setting of Merton's investment problem in an incomplete market and derive its data-driven solutions. More precisely, we demonstrate that, in spite of having no informational benefit, RL can still be used to learn optimal portfolio policies in a model-free manner by employing randomized actions. We propose an auxiliary relaxed control problem with a special class of Gaussian policies within the continuous-time RL exploratory framework developed by Wang et al. (2020) and show that the optimal solution of this auxiliary problem gives rise to that of the original Merton problem. A key insight is that exploration–exploitation tradeoff in the current setting of a small investor is not about information gains versus payoff losses, but about the strength of the learning signals (the gradient estimates of the objective function) versus their reliability (the variance of the gradient estimates). It goes without saying that the RL approach can be extended readily to the problem with a large investor in which the RL will play both the conceptual and technical roles. As such, we believe that the paper resolves the long-standing question about the necessity and applicability of RL in portfolio choice.

We develop an actor–critic RL algorithm for learning optimal policies and value functions iteratively. Through policy evaluation and policy update, we show such an iterative procedure yields monotonically improving policies. Using a stochastic volatility environment as an example, we explain why the traditional model-based, plug-in methods may fail due to sensitivity to model estimation errors. By contrast, the RL methods are model-free and learn optimal policies directly from data, which is naturally robust to the said estimation errors. Numerical results based on both synthetic and market data forcefully demonstrate the efficiency and robustness of our methods against traditional plug-in methods.

This paper brings about many open research questions. A fascinating one is to fully understand the general interactions between the randomness injected by stochastic policies and the randomness in the market, as well as their joint impacts on learning performance.

Acknowledgment

Dai acknowledges the supports of Hong Kong GRF (15213422, 15217123), The Hong Kong Polytechnic University Research Grants (P0039114, P0042456, P0042708, and P0045342), and NSFC (12071333).

References

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

Andrei D, Hasler M (2015) Investor attention and stock market volatility. *The Review of Financial Studies* 28(1):33–72.

Bachouch A, Huré C, Langrené N, Pham H (2021) Deep neural networks algorithms for stochastic control problems on finite horizon: Numerical applications. *Methodology and Computing in Applied Probability* 1–36.

Barndorff-Nielsen OE, Shephard N (2002) Estimating quadratic variation using realized variance. *Journal of Applied Econometrics* 17(5):457–477.

Buehler H, Gonon L, Teichmann J, Wood B (2019) Deep hedging. *Quantitative Finance* 19(8):1271–1291.

Carr P, Geman H, Madan DB, Yor M (2005) Pricing options on realized variance. *Finance and Stochastics* 9:453–475.

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.

Cvitanić J, Lazrak A, Martellini L, Zapatero F (2006) Dynamic portfolio choice with parameter uncertainty and the economic value of analysts' recommendations. *The Review of Financial Studies* 19(4):1113–1156.

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

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

Drimus GG (2012) Options on realized variance by transform methods: A non-affine stochastic volatility model. *Quantitative Finance* 12(11):1679–1694.

Duarte V, Duarte D, Silva DH (2024) Machine learning for continuous-time finance. *The Review of Financial Studies* 37(11):3217–3271.

Epstein LG, Schneider M (2007) Learning under ambiguity. *The Review of Economic Studies* 74(4):1275–1303.

Gao X, Chan L (2000) An algorithm for trading and portfolio management using Q-learning and Sharpe ratio maximization. *Proceedings of the International Conference on Neural Information Processing*, 832–837.

Gennotte G (1986) Optimal portfolio choice under incomplete information. *The Journal of Finance* 41(3):733–746.

Green PJ, Łatuszyński K, Pereyra M, Robert CP (2015) Bayesian computation: A summary of the current state, and samples backwards and forwards. *Statistics and Computing* 25:835–862.

Guijarro-Ordonez J, Pelger M, Zanotti G (2021) Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028* .

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

Han J, E W (2016) Deep learning approximation for stochastic control problems. *arXiv preprint arXiv:1611.07422* .

Hansen LP, Sargent TJ (2001) Robust control and model uncertainty. *American Economic Review* 91(2):60–66.

Hansen LP, Sargent TJ, Turmuhambetova G, Williams N (2006) Robust control and model misspecification. *Journal of Economic Theory* 128(1):45–90.

Hansen LP, Singleton KJ (1982) Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 1269–1286.

Hansen PR, Lunde A (2006) Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24(2):127–161.

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

Huang Y, Jia Y, Zhou X (2022) Achieving mean–variance efficiency by continuous-time reinforcement learning. *Proceedings of the Third ACM International Conference on AI in Finance*, 377–385.

Huang Y, Jia Y, Zhou XY (2024) Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. *Available at SSRN: https://ssrn.com/abstract=5048272* .

Jia Y, Ouyang D, Zhang Y (2025) Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *arXiv preprint arXiv:2503.09981* .

Jia Y, Zhou XY (2022a) Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research* 23(154):1–55.

Jia Y, Zhou XY (2022b) Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research* 23(154):1–55.

Jia Y, Zhou XY (2023) q-Learning in continuous time. *Journal of Machine Learning Research* 24(161):1–61.

Jiang R, Saunders D, Weng C (2022) The reinforcement learning Kelly strategy. *Quantitative Finance* 22(8):1445–1464.

Jin O, El-Saawy H (2016) Portfolio management using reinforcement learning. *Technical Report, Stanford University* .

Kraft H (2005) Optimal portfolios and Heston's stochastic volatility model: An explicit solution for power utility. *Quantitative Finance* 5(3):303–313.

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.

Luenberger DG (1998) *Investment Science* (Oxford University Press: New York).

Maenhout PJ (2004) Robust portfolio rules and asset pricing. *The Review of Financial Studies* 17(4):951–983.

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

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

Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* 8(4):323–361.

Pástor L (2000) Portfolio selection and asset pricing models. *The Journal of Finance* 55(1):179–223.

Reppen AM, Soner HM (2023) Deep empirical risk minimization in finance: Looking into the future. *Mathematical Finance* 33(1):116–145.

Sutton RS, Barto AG (2011) *Reinforcement learning: An Introduction* (Cambridge, MA: MIT Press).

Swait J, Marley AA (2013) Probabilistic choice (models) as a result of balancing multiple goals. *Journal of Mathematical Psychology* 57(1-2):1–14.

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 4378950* .

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.

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

Electronic Companion to "Data-Driven Merton's Strategies via Policy Randomization"

Appendix A: Motivation of Formulation (5)

We explain the exploratory formulation (5) by starting with a discrete-time setting for easy understanding. Divide the whole time interval $[0,T]$ into small intervals of size Δt. Let $R_t := \log W_t$ be the wealth log-return. Given an action $a \in \mathbb{R}$, the instantaneous change of the log return process in the interval $[t, t+\Delta t]$ is

$$\Delta R_t = \left[r + (\mu_t - r)a - \frac{1}{2}\sigma_t^2 a^2 \right] \Delta t + \sigma_t a \Delta B_t.$$

Now, we assume that the agent takes action randomly according to a policy distribution π_t that is independent of the underlying Brownian motions in the market. Focusing on the first and second moments of the randomized policy, we replace a with $e_t + v_t\varepsilon_t$, where ε_t is a random variable with zero mean and unit variance independent of B_t and $\tilde{B}_t$,

$$e_t = \int_{\mathbb{R}} a\boldsymbol{\pi}_t(a)\mathrm{d}a, \quad \text{and} \quad v_t = \sqrt{\int_{\mathbb{R}} a^2\boldsymbol{\pi}_t(a)\mathrm{d}a - (\int_{\mathbb{R}} a\boldsymbol{\pi}_t(a)\mathrm{d}a)^2}.$$

It follows

$$\begin{aligned}\Delta R_t &= \left[r + (\mu_t - r)(e_t + v_t\varepsilon_t) - \frac{1}{2}\sigma_t^2(e_t + v_t\varepsilon_t)^2 \right] \Delta t + \sigma_t(e_t + v_t\varepsilon_t)\Delta B_t \\ &= \left[r + (\mu_t - r)e_t - \frac{1}{2}\sigma_t^2(e_t^2 + v_t^2) \right] \Delta t + \sigma_t e_t \Delta B_t + \sigma_t v_t \varepsilon_t \Delta B_t + \mathrm{Residual}_t,\end{aligned}$$

where the residual term $\mathrm{Residual}_t$ is given as follows:

$$\mathrm{Residual}_t = (\mu_t - r)v_t\varepsilon_t\Delta t - \sigma_t^2 e_t v_t \varepsilon_t \Delta t - \sigma_t^2 v_t^2(\varepsilon_t^2 - 1)\Delta t.$$

Since the residual term is a mean zero random variable of size $O(\Delta t)$ and the strategy noises ε_t are mutually independent between time intervals, by the law of large numbers, the residual term will vanish when we take the sum over the whole time interval and send Δt to zero. In addition, as $\varepsilon_t \Delta B_t$ is a mean zero random variable of size $O(\sqrt{\Delta t})$, its summation is asymptotically Gaussian by the central limit theorem. Furthermore, we have $\mathrm{Cov}(\varepsilon_t\Delta B_t, \Delta B_t) = 0$ and $\mathrm{Cov}(\varepsilon_t \Delta B_t, \Delta \tilde{B}_t) = 0$ as ε_t is independent of B_t and $\tilde{B}_t$. Thus, $\varepsilon_t \Delta B_t$ can be approximately treated as the increment of another Brownian motion independent of B_t and $\tilde{B}_t$. It is not hard to verify that

$$\begin{aligned}\mathbb{E}\left[\Delta R_t\right] &= \left[r + (\mu_t - r)e_t - \frac{1}{2}\sigma_t^2(e_t^2 + v_t^2) \right] \Delta t \\ &= \left[r + (\mu - r)\int_{\mathbb{R}} u\boldsymbol{\pi}_t(u)\mathrm{d}u - \frac{1}{2}\sigma_t^2 \int_{\mathbb{R}} u^2\boldsymbol{\pi}_t(u)\mathrm{d}u \right] \Delta t, \\ \mathrm{Var}\left[\Delta R_t\right] &= \sigma_t^2(e_t^2 + v_t^2)\Delta t + o(\Delta t) = \sigma_t^2 \int_{\mathbb{R}} u^2\boldsymbol{\pi}_t(u)\mathrm{d}u\Delta t + o(\Delta t), \\ \mathrm{Cov}\left[\Delta R_t, \Delta X_t\right] &= \rho\nu\sigma_t e_t \Delta t + o(\Delta t) = \rho\nu\sigma_t \int_{\mathbb{R}} u\boldsymbol{\pi}_t(u)\mathrm{d}u\Delta t + o(\Delta t).\end{aligned}$$

This suggests that at the continuous-time limit, R satisfies the following SDE

$$\begin{aligned}\mathrm{d}R_t = &\left[r + (\mu_t - r)\,\mathrm{Mean}(\boldsymbol{\pi}_t) - \tfrac{1}{2}\sigma_t^2(\mathrm{Mean}(\boldsymbol{\pi}_t))^2 - \tfrac{1}{2}\sigma_t^2\,\mathrm{Var}(\boldsymbol{\pi}_t) \right] \mathrm{d}t \\ &+ \sigma_t \left[\mathrm{Mean}(\boldsymbol{\pi}_t)\mathrm{d}B_t + \sqrt{\mathrm{Var}(\boldsymbol{\pi}_t)}\mathrm{d}\bar{B}_t \right], \quad R_0 = \log w_0,\end{aligned}$$

where $\bar{B}_t$ is another Brownian motion that is mutually independent of B_t and $\tilde{B}_t$. As discussed earlier, $\bar{B}_t$ is introduced to model the additional noise caused by policy randomization and can be regarded as a "random number generator" to generate a randomized policy. The coefficient of the $\mathrm{d}\bar{B}_t$ term involves the variance of $\boldsymbol{\pi}_t$, measuring how much additional noise is introduced into the system.

Applying Itô's formula to the above equation we get that $W_t^{\boldsymbol{\pi}} = e^{R_t}$ satisfies the exploratory dynamics (5). As indicated by the above analysis, this exploratory formulation captures the information up to the second order. Jia et al. (2025) provide a rigorous proof of how the wealth processes under portfolios time-discretely sampled from $\boldsymbol{\pi}$ converge weakly to the solution of (5) when the time step goes to 0.

Appendix B: Learning via Empirical Risk Minimization (ERM)

We document an alternative popular data-driven approach to portfolio problems: optimizing policies through ERM (see, e.g., Reppen and Soner 2023). Specifically, one parameterizes the portfolio policy by a deterministic and sufficiently smooth function of the factor: $a_t = \boldsymbol{u}^\theta(t, X_t)$, and rewrite the wealth equation (3) as

$$\mathrm{d}\log W_t^\theta = \boldsymbol{u}^\theta(t,X_t)\frac{\mathrm{d}S_t}{S_t} + [1-\boldsymbol{u}^\theta(t,X_t)]r\mathrm{d}t - \frac{1}{2}\left(\boldsymbol{u}^\theta(t,X_t)\right)^2 \mathrm{d}\langle\log S\rangle_t.$$

The derivative of $\log W_T^\theta$ in the parameter θ is

$$\frac{\partial \log W_T^\theta}{\partial\theta} = \int_0^T \frac{\partial \boldsymbol{u}^\theta}{\partial\theta}(t,X_t)\left(\frac{\mathrm{d}S_t}{S_t} - r\mathrm{d}t\right) - \left(\boldsymbol{u}^\theta\frac{\partial \boldsymbol{u}^\theta}{\partial\theta}\right)(t,X_t)\mathrm{d}\langle\log S\rangle_t.$$

Therefore, the derivative of the objective in θ is

$$\begin{aligned}&\frac{\partial}{\partial\theta}\mathbb{E}\left[U\left(W_T^\theta\right)\right] = \frac{\partial}{\partial\theta}\mathbb{E}\left[\frac{\exp\{(1-\gamma)\log W_T^\theta\}-1}{1-\gamma}\right]\\ =&\mathbb{E}\left[\left(W_T^\theta\right)^{1-\gamma}\frac{\partial\log W_T^\theta}{\partial\theta}\right] = \mathbb{E}\left[\left(W_T^\theta\right)^{1-\gamma}\int_0^T\frac{\partial\boldsymbol{u}^\theta}{\partial\theta}(t,X_t)\left(\frac{\mathrm{d}S_t}{S_t}-r\mathrm{d}t\right) - \left(\boldsymbol{u}^\theta\frac{\partial\boldsymbol{u}^\theta}{\partial\theta}\right)(t,X_t)\mathrm{d}\langle\log S\rangle_t\right].\end{aligned}$$

One then updates θ using a gradient-based algorithm. Note here we can calculate the gradient of the objective function due to the special structure of the wealth equation (3) along with the assumption that the stock price S_t and the market factor X_t are both exogenous. In a more general setting, computing the gradient may require the knowledge of model primitives. In addition, ERM cannot be applied in real-time (i.e. online) because it requires the observation of the whole stock–factor–wealth process until T. To compare ERM with ours, we consider the Black–Scholes market with the same setting as in Section 3.2.2 where there is no market factor. In this case, $\boldsymbol{u}^\theta$ degenerates into a scalar θ. While updating θ, we apply the same projection and learning rates in our proposed methods in Section 3.2.2. Figure 4 shows the result in terms of ERWL. It is seen that with a small size dataset (i.e. fewer than 100 episodes) ERM performs poorly compared with our methods using $\lambda = 0.01, 0.1, 1$. Only when the dataset size is large is its performance comparable to our method with $\lambda = 1$ and the convergence rate is similar to ours.

Appendix C: Additional Numerical Results

In the main paper, we generate a training dataset with a length of 20 years, and each time, we sample a subsequence with a length of 1 year as one episode for training. This is to capture the practical situation in which financial data are always limited. However, in a simulation study, we can generate as much data as we desire. Here, we report the results of such a "thought experiment" when data are unlimited. Specifically, in

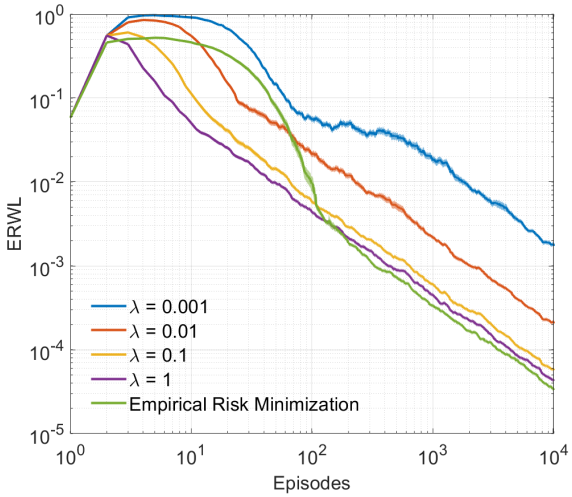

Figure 4 **The comparison between the empirical risk minimization and the proposed method. The horizontal (the number of episodes) and vertical (expected relative wealth loss) axes are both in log-scale. The shaded areas indicate the standard deviations of the estimated ERWLs. The results are based on 1000 times of independent simulation runs and 10,000 episodes of 1-year trajectory is used in each run. The model parameters are $\mu = 0.2, r = 0.02, \sigma = 0.3, \gamma = 3, T = 1$. The learning rate is $a_n = 10/(n+1)$ and the initial policy parameter is $\theta_0 = 0$. The projected region is taken as $c_n = \max\{10, \sqrt{\log(n+1)}\}$ and discretization size is $\Delta t_n = \min\{0.001, 10/(n+1)\}$.**

each episode, we generate *independent* one-year data from the given dynamics for training. Figure 5 illustrates the learning curves of the two RL methods, where average utilities are computed on an independent, fixed test set with 10000 wealth trajectories. Both curves, based on specific forms and neural networks, converge to the omniscient optimal utility after about 3000 independent episodes.

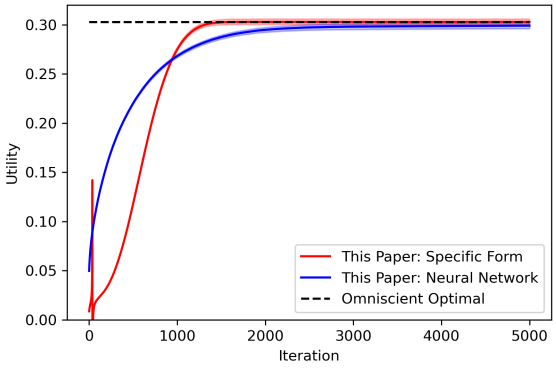

Figure 5 **Average utility of RL algorithms on the test set as functions of the number of training episodes. In each episode, independent one-year data are generated from the model dynamics for training. The width of the shaded area is twice the standard deviation.**

Appendix D: Proof of Statements

D.1. Proof of Theorem 1

We first verify that the function $V^{(\lambda)}$ given in (10) solves the HJB equation (8). Note that $V^{(\lambda)}_{ww} = -\gamma w^{1-\gamma}\exp\{\varphi(t,x)-\lambda(1-\gamma)(T-t)/2\}<0$, hence the "sup" in (8) is achieved at

$$\boldsymbol{u}^*(t,x,w)=\frac{(\mu(t,x)-r)V^{(\lambda)}_w+\rho\nu(t,x)\sigma(t,x)V^{(\lambda)}_{wx}}{-\sigma^2(t,x)wV^{(\lambda)}_{ww}}=\frac{\mu(t,x)-r}{\gamma\sigma^2(t,x)}+\frac{\rho\nu(t,x)}{\gamma\sigma(t,x)}\varphi_x(t,x).$$

It is then straightforward to verify that $V^{(\lambda)}$ satisfies (8).

The rest of the results can be proved following a standard verification approach. We include the proof for reader's convenience.

We first show that for any admissible policy $\boldsymbol{\pi}^{(\lambda)}$, the associated value function $J^{(\boldsymbol{\pi}^{(\lambda)})}$ defined in (6) is smaller than $V^{(\lambda)}$. Let $(W^{\boldsymbol{\pi}^{(\lambda)}},X)$ be the wealth–factor process under $\boldsymbol{\pi}^{(\lambda)}$. Apply Itô's lemma to $V^{(\lambda)}(t,W_t^{\boldsymbol{\pi}^{(\lambda)}},X_t)$ to obtain

$$\begin{aligned}
&V^{(\lambda)}(T,W_T^{\boldsymbol{\pi}^{(\lambda)}},X_T)-V^{(\lambda)}(t,W_t^{\boldsymbol{\pi}^{(\lambda)}},X_t)\\
=&\int_t^T \mathrm{d}s\Bigg\{\frac{\partial V^{(\lambda)}}{\partial t}+[r+(\mu(s,X_s)-r)u_s]W_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_w+\frac{1}{2}\sigma^2(s,X_s)\left(u_s^2+\frac{\lambda}{\gamma\sigma^2(s,X_s)}\right)W_s^{\boldsymbol{\pi}^{(\lambda)}2}V^{(\lambda)}_{ww}\\
&\qquad+m(s,X_s)V^{(\lambda)}_x+\frac{1}{2}\nu^2(s,X_s)V^{(\lambda)}_{xx}+\rho\nu(s,X_s)\sigma(s,X_s)u_sW_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_{wx}\Bigg\}\\
&+\int_t^T\left\{\sigma(s,X_s)u_sW_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_w\mathrm{d}B_t+\sqrt{\frac{\lambda}{\gamma}}W_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_w\mathrm{d}\bar{B}_t+\nu(s,X_s)V^{(\lambda)}_x[\rho\mathrm{d}B_t+\sqrt{1-\rho^2}\mathrm{d}\tilde{B}_t]\right\},
\end{aligned}$$

where $u_s=\mathrm{Mean}(\boldsymbol{\pi}^{(\lambda)}(\cdot|s,W_s^{\boldsymbol{\pi}^{(\lambda)}},X_s))$. Define a sequence of increasing stopping times $\tau_n=\inf\{s\geqslant t:|X_s|\geqslant n$ or $\left(W_s^{\boldsymbol{\pi}^{(\lambda)}}\right)^{1-\gamma}\geqslant n\}$. Replacing T by $T\wedge\tau_n$ in the above formula and taking conditional expectation on both sides, we obtain

$$\begin{aligned}
&\mathbb{E}\left[V^{(\lambda)}(T\wedge\tau_n,W_{T\wedge\tau_n}^{\boldsymbol{\pi}^{(\lambda)}},X_{T\wedge\tau_n})\,|\,W_t^{\boldsymbol{\pi}^{(\lambda)}}=w,X_t=x\right]-V^{(\lambda)}(t,w,x)\\
=&\mathbb{E}\Bigg[\int_t^T\mathrm{d}s\Bigg\{\frac{\partial V^{(\lambda)}}{\partial t}+[r+(\mu(s,X_s)-r)u_s]W_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_w+\frac{1}{2}\sigma^2(s,X_s)\left(u_s^2+\frac{\lambda}{\gamma\sigma^2(s,X_s)}\right)W_s^{\boldsymbol{\pi}^{(\lambda)}2}V^{(\lambda)}_{ww}\\
&\qquad+m(s,X_s)V^{(\lambda)}_x+\frac{1}{2}\nu^2(s,X_s)V^{(\lambda)}_{xx}+\rho\nu(s,X_s)\sigma(s,X_s)u_sW_s^{\boldsymbol{\pi}^{(\lambda)}}V^{(\lambda)}_{wx}\Bigg\}\,|\,W_t^{\boldsymbol{\pi}^{(\lambda)}}=w,X_t=x\Bigg]\\
\leqslant&0,
\end{aligned}$$

where the last inequality is due to the fact that $V^{(\lambda)}$ solves the HJB equation (8), and the $\mathrm{d}B_t,\mathrm{d}\tilde{B}_t$ terms vanish because they are martingales. Moreover,

$$\begin{aligned}
&\mathbb{E}\left[V^{(\lambda)}(T\wedge\tau_n,W_{T\wedge\tau_n}^{\boldsymbol{\pi}^{(\lambda)}},X_{T\wedge\tau_n})\,|\,W_t^{\boldsymbol{\pi}^{(\lambda)}}=w,X_t=x\right]\\
=&-\frac{1}{1-\gamma}+\mathbb{E}\left[\frac{\left(W_T^{\boldsymbol{\pi}^{(\lambda)}}\right)^{1-\gamma}}{1-\gamma}\mathbb{1}_{\{\tau_n>T\}}\,|\,W_t^{\boldsymbol{\pi}^{(\lambda)}}=w,X_t=x\right]\\
&+\mathbb{E}\left[\mathbb{1}_{\{\tau_n\leqslant T\}}\frac{\left(W_{\tau_n}^{\boldsymbol{\pi}^{(\lambda)}}\right)^{1-\gamma}\exp\{\varphi(\tau_n,X_{\tau_n})-\lambda(1-\gamma)(T-\tau_n)/2\}}{1-\gamma}\,|\,W_t^{\boldsymbol{\pi}^{(\lambda)}}=w,X_t=x\right]\\
=:&-\frac{1}{1-\gamma}+\frac{1}{1-\gamma}I_1+\frac{1}{1-\gamma}I_2.
\end{aligned}$$

By the monotone convergence theorem, we have

$$\lim_{n\to\infty} I_1 = \mathbb{E}\left[\left(W_T^{\boldsymbol{\pi}^{(\lambda)}}\right)^{1-\gamma} \mid W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\right].$$

Thus, we have proved

$$V^{(\lambda)}(t,w,x) \geqslant \mathbb{E}\left[U\left(W_T^{\boldsymbol{\pi}^{(\lambda)}}\right) \mid W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\right] + \limsup_{n\to\infty} \frac{1}{1-\gamma} I_2. \tag{29}$$

It suffices to show that $\lim_{n\to\infty} I_2 = 0$. By Hölder's inequality, we have

$$I_2 \leqslant n e^{\lambda T|1-\gamma|} \left(\mathbb{P}(\tau_n \leqslant T)\right)^{1/q} \left(\mathbb{E}\left[\exp\{p\varphi(T\wedge\tau_n, X_{T\wedge\tau_n})\} \mid W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\right]\right)^{1/p},$$

for any $p, q > 1$ satisfying $1/p + 1/q = 1$. By the regularity of the function φ, we have

$$\limsup_{n\to\infty} \mathbb{E}\left[\exp\{p\varphi(T\wedge\tau_n, X_{T\wedge\tau_n})\} \mid W_t^{\boldsymbol{\pi}^{(\lambda)}} = w, X_t = x\right] < \infty.$$

Moreover, standard growth conditions of SDEs yield

$$\mathbb{P}(\tau_n \leqslant T) \leqslant C n^{-L}$$

where L can be arbitrarily small. This implies that $\lim_{n\to\infty} I_2 = 0$.

Next, when the policy (11) is taken, then the inequality (29) becomes an equality, because (11) achieves the supremum in the HJB equation (8) and the policy is admissible based on Definition 1. This establishes the optimality of (11). Finally, the above analysis applies to the case when $\lambda = 0$ noting that $\boldsymbol{u}^*$ is independent of λ. This proves the last statement and completes the proof.

D.2. Proof of Corollary 1

It follows from the form of the optimal value function (10) that

$$V^{(\lambda)}(t,w,x) = V^{(0)}(t, \exp\{\frac{-\lambda(T-t)}{2}\}w, x).$$

The desired result follows from Definition 2.

D.3. Proof of Theorem 2

(i) Given the policy $\boldsymbol{\pi}^{(\lambda)}$, it follows from the Feynman–Kac formula that the value function $J^{(\boldsymbol{\pi}^{(\lambda)})}$ satisfies the linear PDE

$$\begin{aligned}&\frac{\partial J^{(\boldsymbol{\pi}^{(\lambda)})}}{\partial t} + \Big(r + \big(\mu(t,x) - r\big)\boldsymbol{u}(t,x)\Big) w J_w^{(\boldsymbol{\pi}^{(\lambda)})} + \frac{1}{2}\sigma^2(t,x)\Big(\boldsymbol{u}^2(t,x) + \frac{\lambda}{\gamma\sigma(t,x)^2}\Big) w^2 J_{ww}^{(\boldsymbol{\pi}^{(\lambda)})} \\ &+ m(t,x) J_x^{(\boldsymbol{\pi}^{(\lambda)})} + \frac{1}{2}\nu^2(t,x) J_{xx}^{(\boldsymbol{\pi}^{(\lambda)})} + \rho\nu(t,x)\sigma(t,x)\boldsymbol{u}(t,x) w J_{wx}^{(\boldsymbol{\pi}^{(\lambda)})} = 0\end{aligned}$$

with $J^{(\boldsymbol{\pi}^{(\lambda)})}(T,w,x) = U(w)$. A direct calculation verifies that the function $J^{(\boldsymbol{\pi}^{(\lambda)})}$ specified in the statement satisfies the above PDE. The desired result then follows from the uniqueness of the solution to the linear PDE.

(ii) By (i), $J^{(\tilde{\boldsymbol{\pi}}^{(\lambda)})}$ has the same representation (12) with $\bar{\varphi}$ replaced by $\tilde{\varphi}$ while the latter satisfies the PDE (13) with $\boldsymbol{u}$ replaced by $\tilde{\boldsymbol{u}}$. Therefore, it suffices to show that $\tilde{\varphi} \geqslant \bar{\varphi}$ when $0<\gamma<1$, and $\tilde{\varphi} \leqslant \bar{\varphi}$ when $\gamma>1$.

Consider the transformation: $\phi(t,w,x) = e^{\bar{\varphi}(t,w,x)}$. Then, ϕ satisfies the PDE

$$\begin{aligned}&\frac{\partial \phi}{\partial t} + m(t,x)\phi_x + \frac{1}{2}\nu^2(t,x)\phi_{xx}\\ &+ (1-\gamma)\left[r + (\mu(t,x)-r)\boldsymbol{u}(t,x)\phi - \frac{\gamma}{2}\sigma^2(t,x)\boldsymbol{u}^2(t,x)\phi + \rho\nu(t,x)\sigma(t,x)\boldsymbol{u}(t,x)\phi_x\right] = 0.\end{aligned}$$

Similarly, $\tilde{\phi}(t,w,x) = e^{\tilde{\varphi}(t,w,x)}$ satisfies

$$\begin{aligned}&\frac{\partial \tilde{\phi}}{\partial t} + m(t,x)\tilde{\phi}_x + \frac{1}{2}\nu^2(t,x)\tilde{\phi}_{xx}\\ &+ (1-\gamma)\left[r + (\mu(t,x)-r)\tilde{\boldsymbol{u}}(t,x)\tilde{\phi} - \frac{\gamma}{2}\sigma^2(t,x)\tilde{\boldsymbol{u}}^2(t,x)\tilde{\phi} + \rho\nu(t,x)\sigma(t,x)\tilde{\boldsymbol{u}}(t,x)\tilde{\phi}_x\right] = 0,\end{aligned}$$

with $\tilde{\boldsymbol{u}}(t,x) = \frac{(\mu(t,x)-r)\boldsymbol{u}(t,x)\phi(t,x)+\rho\nu(t,x)\sigma(t,x)\boldsymbol{u}(t,x)\phi_x(t,x)}{\gamma\sigma^2(t,x)\phi(t,x)}$. Note that

$$\begin{aligned}&(\mu(t,x)-r)\boldsymbol{u}(t,x)\phi - \frac{\gamma}{2}\sigma^2(t,x)\boldsymbol{u}^2(t,x)\phi + \rho\nu(t,x)\sigma(t,x)\boldsymbol{u}(t,x)\phi_x\\ \leqslant&(\mu(t,x)-r)\tilde{\boldsymbol{u}}(t,x)\phi - \frac{\gamma}{2}\sigma^2(t,x)\tilde{\boldsymbol{u}}^2(t,x)\phi + \rho\nu(t,x)\sigma(t,x)\tilde{\boldsymbol{u}}(t,x)\phi_x.\end{aligned}$$

Therefore, when $0<\gamma<1$, we have

$$\begin{aligned}&\frac{\partial \phi}{\partial t} + m(t,x)\phi_x + \frac{1}{2}\nu^2(t,x)\phi_{xx}\\ &+ (1-\gamma)\left[r + (\mu(t,x)-r)\tilde{\boldsymbol{u}}(t,x)\phi - \frac{\gamma}{2}\sigma^2(t,x)\tilde{\boldsymbol{u}}^2(t,x)\phi + \rho\nu(t,x)\sigma(t,x)\tilde{\boldsymbol{u}}(t,x)\phi_x\right] \geqslant 0.\end{aligned}$$

By the comparison principle of PDEs, we have $\tilde{\phi} \geqslant \phi$ when $0<\gamma<1$. The case for $\gamma>1$ can be proved in parallel. This completes the proof.

D.4. Proof of Theorem 3

(i) The equation in the statement is essentially the martingale orthogonality condition for policy evaluation developed in Jia and Zhou (2022a). Following the same argument as in the proof of Proposition 4 therein, we obtain

$$\hat{V}(t_0,w_0,x_0) = \mathbb{E}\left[U(W_T^{a^{\boldsymbol{\pi}^{(\lambda)}}})\middle| W_{t_0}^{a^{\boldsymbol{\pi}^{(\lambda)}}} = w_0, X_{t_0} = x_0\right],$$

which, by definition, coincides with the value function $J^{(\boldsymbol{\pi}^{(\lambda)})}$.

(ii) Denote $\mu_t = \mu(t,X_t)$, $\sigma_t = \sigma(t,X_t)$, $\eta_t = \eta(t, W_t^{a^{\tilde{\boldsymbol{\pi}}^{(\lambda)}}}, X_t)$, and

$$J_t = J^{(\boldsymbol{\pi}^{(\lambda)})}(t, W_t^{a^{\tilde{\boldsymbol{\pi}}^{(\lambda)}}}, X_t) = \frac{\left(W_t^{a^{\tilde{\boldsymbol{\pi}}^{(\lambda)}}}\right)^{1-\gamma}\exp\{\varphi(t,X_t) - \lambda(1-\gamma)(T-t)/2\} - 1}{1-\gamma},\ \varphi_t = \varphi(t,X_t).$$

Apply Itô's lemma to J_t to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\int_{t_0}^{T}\eta_t\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\mathrm{d}J^{(\boldsymbol{\pi}^{(\lambda)})}(t,W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}},X_t)\right]\\
=&\mathbb{E}\Bigg[\int_{t_0}^{T}\eta_t\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\left(W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}}\right)^{1-\gamma}\exp\{\varphi_t-\lambda(1-\gamma)(T-t)/2\}\times\\
&\qquad\left\{\left[\left(r+(\mu_t-r)a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\frac{\gamma}{2}\sigma_t^2(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}})^2\right)\mathrm{d}t+\cdots\mathrm{d}B_t+\sigma_t a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}\mathrm{d}\langle B,\varphi\rangle_t\right]+\frac{1}{1-\gamma}\left[\mathrm{d}\varphi_t+\frac{1}{2}\mathrm{d}\langle\varphi\rangle_t+\frac{\lambda(1-\gamma)}{2}\mathrm{d}t\right]\right\}\Bigg]\\
=&\mathbb{E}\Bigg[\int_{t_0}^{T}\eta_t\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\left(W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}}\right)^{1-\gamma}\exp\{\varphi_t-\lambda(1-\gamma)(T-t)/2\}\times\\
&\qquad\left\{\left[\left((\mu_t-r)\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)-\gamma\sigma_t^2\hat{\boldsymbol{u}}(t,X_t)\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\right)\mathrm{d}t+\sigma_t\left(a_t^{\hat{\boldsymbol{\pi}}^{(\lambda)}}-\hat{\boldsymbol{u}}(t,X_t)\right)\mathrm{d}\langle B,\varphi\rangle_t\right]\right\}\Bigg]\\
=&\mathbb{E}\left[\int_{t_0}^{T}\eta_t\frac{\lambda}{\gamma\sigma_t^2}\left(W_t^{a^{\hat{\boldsymbol{\pi}}^{(\lambda)}}}\right)^{1-\gamma}\exp\{\varphi_t-\lambda(1-\gamma)(T-t)/2\}\left[(\mu_t-r)\mathrm{d}t-\gamma\sigma_t^2\hat{\boldsymbol{u}}(t,X_t)\mathrm{d}t+\sigma_t\mathrm{d}\langle B,\varphi\rangle_t\right]\right],
\end{aligned}
$$

where $\langle B,\varphi\rangle$ is the covariational process between B_t and φ_t and $\langle\varphi\rangle$ is the quadratic variation process of φ_t. Hence

$$\mathrm{d}\langle B,\varphi\rangle_t=\rho\nu(t,X_t)\varphi_x(t,X_t)\mathrm{d}t.$$

Since the above expectation equals zero for any test process η, the integrand is zero almost surely. Therefore, we have

$$\mu_t-r-\gamma\sigma_t^2\hat{\boldsymbol{u}}(t,X_t)+\rho\sigma_t\nu(t,X_t)\varphi_x(t,X_t)=0$$

or

$$\hat{\boldsymbol{u}}(t,X_t)=\frac{\mu(t,X_t)-r+\rho\nu(t,X_t)\sigma(t,X_t)\varphi_x(t,X_t)}{\gamma\sigma^2(t,X_t)}=\tilde{\boldsymbol{u}}(t,X_t)$$

almost surely and almost all $t\in[t_0,T]$. Because both $\hat{\boldsymbol{u}}$ and $\tilde{\boldsymbol{u}}$ are continuous function, we conclude $\hat{\boldsymbol{u}}(t_0,x_0)=\tilde{\boldsymbol{u}}(t_0,x_0)$. This completes the proof because (t_0,x_0) are arbitrary.

D.5. Proof of Proposition 1

It follows from the wealth equation (3) that

$$\log\left(W_{t_{k+1}}/W_{t_k}\right)=[r+(\mu-r)a_{t_k}-\frac{1}{2}\sigma^2a_{t_k}^2]\Delta t+a_{t_k}\sigma(B_{t_{k+1}}-B_{t_k}).$$

Hence

$$
\begin{aligned}
&\mathbb{E}\left[\left(W_{t_{k+1}}/W_{t_k}\right)^{1-\gamma}|a_{t_k},W_{t_k}\right]\\
=&\exp\left\{(1-\gamma)[r+(\mu-r)a_{t_k}-\frac{1}{2}\sigma^2a_{t_k}^2]\Delta t+\frac{1}{2}(1-\gamma)^2a_{t_k}^2\sigma^2\Delta t\right\}\\
=&\exp\left\{(1-\gamma)\left[r+(\mu-r)\theta-\frac{\gamma}{2}\sigma^2\theta^2+(\mu-r-\theta\gamma\sigma^2)(a_{t_k}-\theta)-\frac{\gamma}{2}\sigma^2(a_{t_k}-\theta)^2\right]\Delta t\right\}.
\end{aligned}
$$

Denote

$$
\begin{aligned}
&A_1=(1-\gamma)\left[r+(\mu-r)\theta-\frac{\gamma}{2}\sigma^2\theta^2+\frac{1}{2}\lambda\right]-\psi,\ A_2=(1-\gamma)(\mu-r-\theta\gamma\sigma^2)\sqrt{\frac{\lambda}{\gamma\sigma^2}}\\
&A_3=(\gamma-1)\lambda/2,\ A_4=2(1-\gamma)\left[r+(\mu-r)\theta-\frac{2\gamma-1}{2}\sigma^2\theta^2+\frac{1}{2}\lambda\right]-2\psi\\
&A_5=2(1-\gamma)(\mu-r-\frac{2\gamma-1}{2}\theta\sigma^2)\sqrt{\frac{\lambda}{\gamma\sigma^2}},\ A_6=(\gamma-1)(2\gamma-1)\lambda/\gamma.
\end{aligned}
$$

Let $L_k = \left(W_{t_{k+1}}/W_{t_k}\right)^{1-\gamma} \exp\left\{\left[-\psi + \lambda(1-\gamma)/2\right]\Delta t\right\}$. Then

$$\begin{aligned}
&\mathbb{E}\left[L_k - 1\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(W_{t_{k+1}}/W_{t_k}\right)^{1-\gamma} \exp\left\{\left[-\psi + \lambda(1-\gamma)/2\right]\Delta t\right\} - 1 \middle| a_{t_k}, W_{t_k}\right]\right] \\
=&\mathbb{E}\left[\exp\left\{(1-\gamma)\left[(\mu - r - \theta\gamma\sigma^2)(a_{t_k} - \theta) - \frac{\gamma}{2}\sigma^2(a_{t_k} - \theta)^2\right]\Delta t\right\} e^{A_1\Delta t} - 1\right] \\
=&e^{A_1\Delta t}\frac{\exp\left\{\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{\sqrt{1-\lambda(\gamma-1)\Delta t}} - 1
\end{aligned}$$

and, therefore,

$$\begin{aligned}
&\mathbb{E}\left[(a_{t_k} - \theta)(L_k - 1)\right] \\
=&\mathbb{E}\left[(a_{t_k} - \theta)\mathbb{E}\left[\left(W_{t_{k+1}}/W_{t_k}\right)^{1-\gamma} \exp\left\{\left[-\psi + \lambda(1-\gamma)/2\right]\Delta t\right\} - 1 \middle| a_{t_k}, W_{t_k}\right]\right] \\
=&\mathbb{E}\left[(a_{t_k} - \theta)\left(\exp\left\{(1-\gamma)\left[(\mu - r - \theta\gamma\sigma^2)(a_{t_k} - \theta) - \frac{\gamma}{2}\sigma^2(a_{t_k} - \theta)^2\right]\Delta t\right\} e^{A_1\Delta t} - 1\right)\right] \\
=&\sqrt{\frac{\lambda}{\gamma\sigma^2}} e^{A_1\Delta t} A_2\Delta t \frac{\exp\left\{\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}}.
\end{aligned}$$

Similarly, we obtain

$$\mathbb{E}\left[(a_{t_k} - \theta)^2 L_k\right] = \frac{\lambda}{\gamma\sigma^2} e^{A_1\Delta t}\left[1 + A_2^2(\Delta t)^2 - 2A_3\Delta t\right]\frac{\exp\left\{\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{5/2}}.$$

$$\mathbb{E}\left[(a_{t_k} - \theta)^2 L_k^2\right] = \frac{\lambda}{\gamma\sigma^2} e^{A_4\Delta t}\left[1 + A_5^2(\Delta t)^2 - 2A_6\Delta t\right]\frac{\exp\left\{\frac{A_5^2(\Delta t)^2}{2[1-A_6\Delta t]}\right\}}{(1-A_6\Delta t)^{5/2}}.$$

Now we can compute

$$\begin{aligned}
&\mathbb{E}\left[\widehat{e(\psi,\theta)}\right] \\
=&\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\gamma\sigma^2(a_{t_k} - \theta)}{\lambda(1-\gamma)}\left[\left(\frac{W_{t_{k+1}}}{W_{t_k}}\right)^{1-\gamma}\exp\left\{\left[-\psi + \lambda(1-\gamma)/2\right]\Delta t\right\} - 1\right]\right] \\
=&\frac{\gamma\sigma^2 K}{\lambda(1-\gamma)}\mathbb{E}\left[(a_{t_k} - \theta)\left(L_k - 1\right)\right] \\
=&\frac{\gamma\sigma^2 K}{\lambda(1-\gamma)}\sqrt{\frac{\lambda}{\gamma\sigma^2}} e^{A_1\Delta t} A_2\Delta t \frac{\exp\left\{\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}} \\
=&T(\mu - r - \theta\gamma\sigma^2)\frac{\exp\left\{A_1\Delta t + \frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}}.
\end{aligned}$$

Under the conditions provided on Δt, there exists a constant C that only depends on μ, r, σ, T such that $|A_1\Delta t| \leqslant C(\mu, r, \sigma, \gamma, T)$, $A_2^2(\Delta t)^2 \leqslant C(\mu, r, \sigma, \gamma, T)$, and $1 - A_3\Delta t \in (\frac{3}{4}, \frac{5}{4})$. Thus,

$$\frac{\exp\left\{A_1\Delta t + \frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}} - 1 \geqslant \frac{1}{(1-A_3\Delta t)^{3/2}}\left(A_1\Delta t + \frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right) \geqslant -C|A_1|\Delta t,$$

$$\frac{\exp\left\{A_1\Delta t + \frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}} - 1 \leqslant \left(1 + C|A_1|\Delta t + CA_2^2(\Delta t)^2\right)(1 + C|A_3|\Delta t) - 1 \leqslant C(|A_1| + |A_3|)\Delta t.$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\widehat{e(\psi,\theta)}\right]-T(\mu-r-\theta\gamma\sigma^2)\right| \\
\leqslant& T(\mu-r-\theta\gamma\sigma^2)\left|\frac{\exp\left\{A_1\Delta t+\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{3/2}}-1\right| \\
\leqslant& C(1+|\theta^2|+|\psi|+\lambda)\Delta t.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathrm{Var}\left[\widehat{e(\psi,\theta)}\right] \\
=&\sum_{k=0}^{K-1}\mathrm{Var}\left[\frac{\gamma\sigma^2(a_{t_k}-\theta)}{\lambda(1-\gamma)}\left(L_k-1\right)\right]=\left(\frac{\gamma\sigma^2}{\lambda(1-\gamma)}\right)^2 K\,\mathrm{Var}\left[\left(a_{t_k}-\theta\right)\left(L_k-1\right)\right] \\
\leqslant& K\left(\frac{\gamma\sigma^2}{\lambda(1-\gamma)}\right)^2\mathbb{E}\left[(a_{t_k}-\theta)^2(L_k-1)^2\right] \\
\leqslant& K\left(\frac{\gamma\sigma^2}{\lambda(1-\gamma)}\right)^2\frac{\lambda}{\gamma\sigma^2}\Bigg[e^{A_4\Delta t}\left[1+A_5^2(\Delta t)^2-4A_6\Delta t\right]\frac{\exp\left\{\frac{A_5^2(\Delta t)^2}{2[1-A_6\Delta t]}\right\}}{(1-A_6\Delta t)^{5/2}} \\
&-2e^{A_1\Delta t}\left[1+A_2^2(\Delta t)^2-2A_3\Delta t\right]\frac{\exp\left\{\frac{A_2^2(\Delta t)^2}{2[1-A_3\Delta t]}\right\}}{(1-A_3\Delta t)^{5/2}}+1\Bigg] \\
\leqslant& K\frac{\gamma\sigma^2}{\lambda(1-\gamma)^2}\left[(A_4+\frac{1}{2}A_6-2A_1-A_3)\Delta t+C(1+\psi^2+\theta^4+\lambda^2)(\Delta t)^2\right] \\
\leqslant& C(1+\frac{\theta^2}{\lambda})+C\left(\frac{1+\psi^2+\theta^4}{\lambda}+\lambda\right)\Delta t.
\end{aligned}$$

D.6. Proof of Theorem 4

We first prove a result regarding equivalent relative wealth loss (ERWL) in the Black–Scholes market.

LEMMA 2. *In the Black–Scholes market, the equivalent relative wealth loss of a determinist policy* $\boldsymbol{u}^\theta\equiv\theta$ *is*

$$\mathrm{ERWL}(\boldsymbol{u}^\theta)=1-\exp\{-\frac{T\gamma\sigma^2}{2}(\theta-\theta^*)^2\}\leqslant\frac{T\gamma\sigma^2}{2}(\theta-\theta^*)^2,$$

where $\theta^*=\frac{\mu-r}{\gamma\sigma^2}$ *is the ground truth optimal allocation.*

Proof of Lemma 2 We show by direct calculation. Under the deterministic policy $\boldsymbol{u}^\theta=\theta$, it follows from (3) that the corresponding wealth process W^θ satisfies

$$\frac{\mathrm{d}W_t^\theta}{W_t^\theta}=\theta\frac{\mathrm{d}S_t}{S_t}+(1-\theta)r\mathrm{d}t=[r+(\mu-r)\theta]\mathrm{d}t+\sigma\theta\mathrm{d}B_t.$$

Hence, $\log W_T^\theta\sim\mathcal{N}\left(\log w_0+[r+(\mu-r)\theta-\frac{1}{2}\sigma^2\theta^2]T,\sigma^2\theta^2T\right)$, leading to

$$\begin{aligned}
J(0,w_0)=&\frac{1}{1-\gamma}\mathbb{E}\left[e^{(1-\gamma)\log W_T^\theta}\right]-\frac{1}{1-\gamma} \\
=&\frac{w_0^{1-\gamma}}{1-\gamma}\exp\{(1-\gamma)[r+(\mu-r)\theta-\frac{1}{2}\sigma^2\theta^2]T+\frac{(1-\gamma)^2}{2}\sigma^2\theta^2T\}-\frac{1}{1-\gamma} \\
=&\frac{w_0^{1-\gamma}\exp\{(1-\gamma)T[r+(\mu-r)\theta-\frac{\gamma}{2}\sigma^2\theta^2]\}-1}{1-\gamma} \\
=&\frac{w_0^{1-\gamma}\exp\left\{-\frac{(1-\gamma)\gamma T\sigma^2}{2}(\theta-\theta^*)^2+[r+\frac{(\mu-r)^2}{2\gamma\sigma^2}](1-\gamma)T\right\}-1}{1-\gamma} \\
=&V^{(0)}\left(0,w_0e^{-\frac{\gamma T\sigma^2}{2}(\theta-\theta^*)^2}\right).
\end{aligned}$$

Hence, by Definition 2, we have

$$\mathrm{ERWL}(\boldsymbol{u}^\theta) = 1 - \exp\{-\frac{T\gamma\sigma^2}{2}(\theta-\theta^*)^2\} \leqslant \frac{T\gamma\sigma^2}{2}(\theta-\theta^*)^2,$$

where the inequality follows from the basic inequality $e^z \geqslant 1+z$ with $z = -\frac{T\gamma\sigma^2}{2}(\theta-\theta^*)^2$.

The next lemma is about a particular recursive relation.

LEMMA 3. *Suppose $\{e_n\}_{n\geqslant n_0}$ is a sequence of positive real numbers and $n_0 \geqslant 4$ satisfying*

$$e_{n+1} \leqslant (1-\alpha_n)e_n + (C_1 + C_2\log n)\alpha_n^2,\ \forall n \geqslant n_0,$$

where $\{\alpha_n\}_{n\geqslant 0}$ is a positive sequence satisfying $\alpha_n \leqslant \alpha_{n+1}(1+Aa_{n+1})$ for all $n \geqslant n_0$ and some $A \in (0,1)$. Let $C = \frac{1}{1-A}\sup_{n\geqslant n_0}\frac{C_1+C_2\log n}{\log(n-1)} \vee \frac{e_{n_0+1}}{\alpha_{n_0}}$. Then $e_{n+1} \leqslant C\alpha_n \log n\ \forall n \geqslant n_0$.

Proof of Lemma 3 We prove by induction. The conclusion holds for $n = n_0$ because $C \geqslant \frac{e_{n_0+1}}{\alpha_{n_0}}$. Assuming the conclusion holds for all $n \leqslant k$ with $k \geqslant n_0$, we examine the case with $n = k+1$. By the given recursive condition and the induction assumption, we have

$$\begin{aligned}
e_{k+1} \leqslant& (1-\alpha_k)e_k + (C_1+C_2\log k)\alpha_k^2 \leqslant (1-\alpha_k)C\alpha_{k-1}\log(k-1) + (C_1+C_2\log k)\alpha_k^2\\
=& C\alpha_k\log k\left(\alpha_{k-1}\frac{1-\alpha_k}{\alpha_k}\log(k-1) + \alpha_k\frac{C_1+C_2\log k}{C\log k}\right)\\
\leqslant& C\alpha_k\log k\left[(1+A\alpha_k)(1-\alpha_k)\frac{\log(k-1)}{\log k} + \alpha_k\frac{C_1+C_2\log k)}{C\log k}\right]\\
<& C\alpha_k\log k + C\alpha_k^2\log k\left(-A\alpha_k\frac{\log(k-1)}{\log k} - (1-A)\frac{\log(k-1)}{\log k} + \frac{C_1+C_2\log k)}{C\log k}\right)\\
<& C\alpha_k\log k + \alpha_k^2\left[C_1 + C_2\log k - (1-A)C\log(k-1)\right] < C\alpha_k\log k,
\end{aligned}$$

where the last inequality is because $C(1-A) > \sup_{n\geqslant n_0}\frac{C_1+C_2\log n}{\log(n-1)} \geqslant \frac{C_1+C_2\log k}{\log(k-1)}$. This proves the desired result.

We are now ready to prove Theorem 4. By Lemma 2, we only need to focus on estimating $\mathbb{E}\left[(\theta_n-\theta^*)^2\right]$.

Recall that θ_n satisfies the recursion $\theta_{n+1} = \Pi_{K_{n+1}}\left(\theta_n + \ell_n\widehat{e_n(\psi,\theta)}\right)$, where $\widehat{e_n(\psi,\theta)}$ is specified in the statement of Theorem 4. Hence, by Proposition 1, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\widehat{e_n(\psi,\theta)}\middle|\theta_n,\phi_n\right] - h(\theta_n)\right| \leqslant C(1+|\theta_n|+|\psi_n|+\lambda)\Delta t_n =: \beta_n\\
&\mathrm{Var}\left[\widehat{e_n(\psi,\theta)} - h(\theta_n)\middle|\theta_n,\phi_n\right] \leqslant C(1+\frac{|\theta_n|^2}{\lambda}) + C(\frac{1+|\psi_n|^2+|\theta_n|^2}{\lambda}+\lambda)\Delta t_n =: \zeta_n
\end{aligned},$$

where $h(\theta) = T\gamma\sigma^2(\theta^*-\theta)$, and C is a constant that only depends on $\mu, r, \sigma, \gamma, T$.

Write $\widehat{e_n(\psi,\theta)} = h(\theta_n) + \xi_n$, where $\mathbb{E}\left[\xi_n\middle|\theta_n,\phi_n\right] \leqslant \beta_n$ and $\mathrm{Var}\left[\xi_n\middle|\theta_n,\phi_n\right] \leqslant \zeta_n$. By the properties of the projection mapping, we have

$$|\theta_{n+1}-\theta^*|^2 \leqslant |\theta_n - \theta^* + \ell_n\left(h(\theta_n)+\xi_n\right)|^2.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[|\theta_{n+1}-\theta^*|^2\middle|\theta_n,\psi_n\right]\\
\leqslant& (1-\ell_nT\gamma\sigma^2)^2(\theta_n-\theta^*)^2 + 2\ell_n(\theta_n-\theta^*)\beta_n + 2\ell_n^2T\gamma\sigma^2(\theta^*-\theta_n)\beta_n + \ell_n^2(\beta_n^2+\zeta_n)\\
\leqslant& (1-\ell_nT\gamma\sigma^2)^2(\theta_n-\theta^*)^2 + \ell_n(1-\ell_nT\gamma\sigma^2)\left[(\theta_n-\theta^*)^2+1\right]\beta_n + \ell_n^2(\beta_n^2+\zeta_n)\\
=& (1-\ell_nT\gamma\sigma^2)(1-\ell_nT\gamma\sigma^2+\ell_n\beta_n)(\theta_n-\theta^*)^2 + \ell_n(1-\ell_nT\gamma\sigma^2)\beta_n + \ell_n^2(\beta_n^2+\zeta_n).
\end{aligned}$$

By a property of projection, we know $|\theta_n| \leqslant c_n \leqslant \sqrt{\log n}$. Hence, for all $n \geqslant n_0$, we have

$$\beta_n \leqslant C(\lambda + 1 + M + c_n)\Delta t_n \leqslant C(\lambda + 1 + M + \log n)\Delta t_n \wedge 1,$$

and

$$\zeta_n \leqslant C(1 + c_n^2 \lambda^{-1}) + C(\lambda + \frac{1 + M^2 + c_n^2}{\lambda})\Delta t_n \leqslant 2C(1 + \lambda^{-1} \log n) + C\lambda \Delta t_n.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[|\theta_{n+1} - \theta^*|^2 \,|\theta_n, \psi_n\right] \\
\leqslant &(1 - \ell_n T\gamma\sigma^2)(1 - \ell_n T\gamma\sigma^2 + \ell_n \beta_n)(\theta_n - \theta^*)^2 + \ell_n(1 - \ell_n T\gamma\sigma^2)\beta_n + \ell_n^2(\beta_n^2 + \zeta_n) \\
\leqslant &(1 - \ell_n T\gamma\sigma^2)(\theta_n - \theta^*)^2 + \ell_n C(\lambda + 1 + M + \log n)\Delta t_n + \ell_n^2 \left(2C + 2C\lambda^{-1} \log n + C\lambda \Delta t_n\right).
\end{aligned}$$

Taking expectation, we obtain a recursive relation for $\mathbb{E}\left[(\theta_n - \theta^*)^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[(\theta_{n+1} - \theta^*)^2\right] \leqslant &(1 - \ell_n T\gamma\sigma^2)\mathbb{E}\left[(\theta_n - \theta^*)^2\right] + \ell_n^2 \left(\frac{C(\lambda + 1 + M + \log n)\Delta t_n}{\ell_n} + 2C + 2C\lambda^{-1} \log n + C\lambda \Delta t_n\right) \\
\leqslant &(1 - \ell_n T\gamma\sigma^2)\mathbb{E}\left[(\theta_n - \theta^*)^2\right] + \ell_n^2 C_2(1 + \lambda^{-1} \log n),
\end{aligned}$$

where C_2 is a constant that only depends on $\mu, r, \sigma, \gamma, T$.

By the specification of ℓ_n in the condition, a direct calculation verifies that ℓ_n satisfies $\ell_n \leqslant \ell_{n+1}(1 + \eta_2 \ell_{n+1})$, for all $n \geqslant n_0$. It follows now from Lemma 3 that for all $n \geqslant n_0$,

$$\mathbb{E}\left[(\theta_{n+1} - \theta^*)^2\right] \leqslant C_1 \ell_n \log n,$$

for some C_1 that is independent of n. In particular, C_1 can be taken such that $C_1 > \frac{C_3}{1-\eta_2} \sup_{n \geqslant n_0} \frac{1 + \lambda^{-1} \log n}{\log(n-1)} \approx \frac{C_3 \lambda^{-1}}{1-\eta_2}$. The proof is completed.