Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

Exploratory mean-variance portfolio selection with Choquet regularizers

Junyi Guo, Xia Han & Hao Wang

To cite this article: Junyi Guo, Xia Han & Hao Wang (13 Oct 2025): Exploratory mean-variance portfolio selection with Choquet regularizers, Quantitative Finance, DOI: 10.1080/14697688.2025.2563094

To link to this article: https://doi.org/10.1080/14697688.2025.2563094

Quantitative Finance, 2025
https://doi.org/10.1080/14697688.2025.2563094

Exploratory mean-variance portfolio selection with Choquet regularizers

JUNYI GUO†, XIA HAN‡ and HAO WANG§*

†School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China
‡School of Mathematical Sciences, LPMC and AAIS, Nankai University, Tianjin 300071, People's Republic of China
§School of Mathematical Sciences, Nankai University, Tianjin 300071, People's Republic of China

(*Received 14 March 2024; accepted 11 September 2025*)

In this paper, we study a continuous-time exploratory mean-variance (EMV) problem under the framework of reinforcement learning (RL), and the Choquet regularizers are used to measure the level of exploration. By applying the classical Bellman principle of optimality, the Hamilton-Jacobi-Bellman equation of the EMV problem is derived and solved explicitly via maximizing statically a mean-variance constrained Choquet regularizer. In particular, the optimal distributions form a location-scale family, whose shape depends on the choices of the Choquet regularizer. We further reformulate the continuous-time Choquet-regularized EMV problem using a variant of the Choquet regularizer. Several examples are given under specific Choquet regularizers that generate broadly used exploratory samplers such as exponential, uniform and Gaussian. Finally, we develop a reinforcement learning algorithm and assess its performance via simulations and empirical analysis, including comparisons with the plug-in policy and the entropy-regularized policy.

Keywords: Choquet regularization; Mean-variance problem; Reinforcement learning; Stochastic control

JEL Classifications: C61, C63, G11

1. Introduction

Reinforcement learning (RL) is an active subarea of machine learning. In RL, the agent can directly interact with the black box environment and get feedback. This kind of learning that focuses on the interaction process between the agent and the environment is called trial-and-error learning. By trial and error learning, we skip the parameter estimation of the model and directly learn the optimal policy (Sutton and Barto 2018), which can overcome some difficulties that traditional optimization theory may have in practice. Many RL algorithms are based on traditional deterministic optimization, and the optimal solution is usually a deterministic policy. But in some situations, it makes sense to solve for an optimal stochastic policy for exploration purposes. The stochastic policy is to change the determined action into a probability distribution through randomization. Searching for the optimal stochastic policy has many advantages, such as robustness (Ziebart 2010) and better convergence (Gu *et al.* 2016) when the system dynamics are uncertain.

Entropy measures the randomness of the actions an agent takes, and thus can indicate the level of exploration in RL. The idea of maximum entropy RL is to make the policy more random in addition to maximizing the cumulative reward, so entropy together with a temperature parameter is added to the objective function as a regularization term; see e.g. Neu *et al.* (2017). Here, the temperature parameter is a regularization coefficient used to control the importance of entropy; the larger the parameter, the stronger the exploratory ability, which helps to accelerate the subsequent policy learning and reduces the possibility of the policy converging to a local optimum. Haarnoja *et al.* (2017) generalized maximum entropy RL to continuous state and continuous action settings rather than tabular settings. Wang *et al.* (2020a) first established a continuous-time RL framework with continuous state and action from the perspective of stochastic control and proved that the optimal exploration policy for the linear-quadratic (LQ) control problem in the infinite time horizon is Gaussian. Further, Wang and Zhou (2020) applied this RL framework for the first time to solve the continuous-time mean-variance (MV) problem, and we refer to Zhou (2021) for more summaries. Motivated by Wang *et al.* (2020a), Dai *et al.* (2023) extended the exploratory stochastic control framework to an

*Corresponding author. Email: hao.wang@mail.nankai.edu.cn

incomplete market, where the asset return correlates with a stochastic market state, and learned an equilibrium policy under a mean-variance criterion. Jiang *et al.* (2022) studied the exploratory Kelly problem by considering both the amount of investment in stock and the portion of wealth in stock as the control for a general time-varying temperature parameter.

From the perspective of risk measures, Han *et al.* (2023) first introduced another kind of index that can measure the randomness of actions called Choquet regularizers. They showed that the optimal exploration distribution of LQ control problem with infinite time horizon is no longer necessarily Gaussian as in Wang *et al.* (2020a), but are dictated by the choice of Choquet regularizers. As mentioned in Han *et al.* (2023), Choquet regularizers have a number of theoretical and practical advantages to be used for RL. In particular, they satisfy several 'good' properties such as quantile additivity, normalization, concavity, and consistency with convex order (mean-preserving spreads) that facilitate analysis as regularizers. Moreover, the availability of a large class of Choquet regularizers makes it possible to compare and choose specific regularizers to achieve certain objectives specific to each learning problem. To the best of our knowledge, there is no literature using regularizers other than entropy to quantify the information gain of exploring the environment for practical problems. Thus, it is natural to consider some practical exploratory stochastic control problems using the Choquet regularizers for regularization.

This paper mainly studies the continuous-time exploratory mean-variance (EMV) problem as in Wang and Zhou (2020) in which we replace the differential entropy used for regularization with the Choquet regularizers. When looking for pre-committed optimal strategies as the goal, the MV model can be converted into a LQ model in finite time horizon by Zhou and Li (2000). The form of the LQ-specialized HJB equation suggests that the problem boils down to a static optimization where the given Choquet regularizer is to be maximized over distributions with given mean and variance, which has been solved by Liu *et al.* (2020). Since the EMV portfolio selection is formulated in a finite time horizon, we show that the optimal distributions form a location-scale family with a time-decaying variance whose shape depends on the choice of Choquet regularizers. This suggests that the level of exploration decreases as the time approaches the end of the planning horizon. We further give the optimal exploration strategies under several specific Choquet regularizers, and observe insights of the perfect separation between exploitation and exploration in the mean and variance of the optimal distribution and the positive effect of a random environment on learning. Furthermore, by utilizing policy improvement and convergence theorems, we devise an RL algorithm to tackle EMV problems using the continuous-time policy gradient method introduced by Jia and Zhou (2022b), subsequently validating it through simulation.

We assert that our paper represents more than a mere extension of the works by Wang and Zhou (2020) and Han *et al.* (2023). Han *et al.* (2023) primarily provided theoretical insights into the use of Choquet integrals as the regularization, without practical algorithmic implementation. For the first time, we introduce an RL algorithm designed to learn the solution of the MV problem under Choquet regularization and generate implementable portfolio allocation strategies, without presupposing any knowledge about the underlying parameters. Our numerical simulations demonstrate the practical effectiveness of the RL algorithm based on Choquet regularization, exhibiting comparable performance to the approach proposed in Wang and Zhou (2020). Moreover, unlike the finding in Wang and Zhou (2020) where the optimal distribution is always Gaussian when the entropy is utilized as the regularization, our study benefits from a wide range of Choquet regularizers. This allows for the comparison and selection of specific regularizers tailored to achieve specific learning objectives, thereby accommodating the preferences of individual agents.

Additionally, inspired by the form of entropy, we further refine our approach by reformulating the continuous-time Choquet-regularized RL problem using a new type of Choquet regularizers called logarithmic Choquet regularizers. Thanks to the monotonic nature of the logarithmic function, the problem remains solvable by maximizing the Choquet regularizer over distributions with given mean and variance. Furthermore, since the regularizers affect the value function, it is to be expected that the variance of the optimal distributions is different. We also examine the explicit exploration costs associated with these different types of regularizers and explored their connections with classical and EMV problems. Notably, we observe distinct differences in exploration costs between the two EMV problems. Specifically, with Choquet regularizers, exploration costs depend on unknown model parameters and specific regularizers. This enables the comparison and selection of specific Choquet regularizers tailored to meet the exploration cost preferences of individual agents. In contrast, with logarithmic Choquet regularizers, exploration costs only depend on the exploration parameter and the time horizon, similar to using entropy as the regularizer in Wang and Zhou (2020). It is interesting to note that the optimization problem under entropy regularization in Wang and Zhou (2020) is equivalent to that under logarithmic Choquet regularization for a specific choice of distortion function. This reveals that the logarithmic Choquet regularizers generalize Shannon entropy regularization as a special case, thereby offering a more flexible framework.

The rest of this paper is organized as follows. Section 2 introduces the MV problem under the Choquet regularizations. Section 3 solves the continuous-time EMV problem and gives several examples. Section 4 discusses the corresponding results under the variant of Choquet regularizations. Section 5 introduces the RL algorithm, and section 6 evaluates its performance through simulation studies and real financial data, comparing it with plug-in strategies and the entropy-regularized method of Wang and Zhou (2020). Section 7 concludes the paper.

2. Formulation of problem

2.1. *Choquet regularizers*

We assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is an atomless probability space. With a slight abuse of notation, let $\mathcal{M}$ denote both the set of

(probability) distribution functions of real random variables and the set of Borel probability measures on $\mathbb{R}$, with the obvious identity $\Pi(x) \equiv \Pi((-\infty, x])$ for $x \in \mathbb{R}$ and $\Pi \in \mathcal{M}$. We denote by $\mathcal{M}^p \subset \mathcal{M}$, $p \in [1, \infty)$, the set of distribution functions or probability measures with finite p-th moment. For a random variable X and a distribution Π, we write $X \sim \Pi$ if the distribution of X is Π under $\mathbb{P}$, and $X \overset{\mathrm{d}}{=} Y$ if two random variables X and Y have the same distribution. We denote by μ and σ^2 the mean and variance functionals on $\mathcal{M}^2$, respectively; that is, $\mu(\Pi)$ is the mean of Π and $\sigma^2(\Pi)$ the variance of Π for $\Pi \in \mathcal{M}^2$. We denote by $\mathcal{M}^2(m, s^2)$ the set of $\Pi \in \mathcal{M}^2$ satisfying $\mu(\Pi) = m \in \mathbb{R}$ and $\sigma^2(\Pi) = s^2 > 0$.

In Han *et al.* (2023), the *Choquet regularizer* is defined to measure and manage the level of exploration for RL based on a subclass of signed Choquet integrals (Wang *et al.* 2020b). Given a concave function $h : [0, 1] \to \mathbb{R}$ of bounded variation with $h(0) = h(1) = 0$ and $\Pi \in \mathcal{M}$, the Choquet regularizer Φ_h on $\mathcal{M}$ is defined as

$$\Phi_h(\Pi) = \int_{\mathbb{R}} h \circ \Pi([x, \infty)) \, \mathrm{d}x.$$

The set of all such functions h is denoted by $\mathcal{H}$.

The following proposition summarizes several useful properties of Φ_h that were previously established in Section 2 of Han *et al.* (2023).

PROPOSITION 2.1 *For $h \in \mathcal{H}$, the Choquet regularizer Φ_h satisfies the following properties:*

(i) *Φ_h is well defined, non-negative, and satisfies location invariant and scale homogeneous.*†
(ii) *$\Phi_h(\delta_c) = 0$, $\forall c \in \mathbb{R}$, where δ_c is the Dirac measure at c.*
(iii) *For all $\Pi_1, \Pi_2 \in \mathcal{M}$ and $\lambda \in [0, 1]$, $\Phi_h(\lambda\Pi_1 + (1 - \lambda)\Pi_2) \geqslant \lambda\Phi_h(\Pi_1) + (1 - \lambda)\Phi_h(\Pi_2)$.*
(iv) *$\Phi_h(\Pi_1) \leqslant \Phi_h(\Pi_2)$ for all $\Pi_1, \Pi_2 \in \mathcal{M}$ with $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$.*‡

Each property in Proposition 2.1 has a natural interpretation for a regularizer that measures the level of randomness, or the level of exploration in the RL context of this paper. Proposition 2.1(i) implies that any distribution for exploration can be measured in non-negative values. Moreover, the measure is invariant to shifts in location and scales linearly with the distribution, which aligns with intuitive properties of randomness. Proposition 2.1(ii) means that degenerate distributions do not have any randomness measured by Φ_h. Proposition 2.1(iii) means that mixing distributions generally introduces additional randomness. Moreover, in part (iv), $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$ is equivalent to saying that Π_2 is a mean-preserving spread of Π_1, indicating that Π_2 is more dispersed and therefore 'more random' than Π_1. Taken together, these properties justify the use of Φ_h as a principled regularizer for quantifying randomness or exploration in reinforcement learning.

The family of Choquet regularizers encompasses several classical dispersion measures, including the range, mean-median deviation, Gini deviation, and inter-ES differences; see Section 2.6 of Wang *et al.* (2020b) for details. In contrast to Shannon entropy, Choquet regularization allows for the selection of different h-functions tailored to the agent's preferences, providing greater flexibility and adaptability in measuring randomness.

For a distribution $\Pi \in \mathcal{M}$, let its left-quantile for $p \in (0, 1]$ be defined as

$$Q_\Pi(p) = \inf\{x \in \mathbb{R} : \Pi(x) \geqslant p\}.$$

It is useful to note that Φ_h admits a quantile representation. Specifically, if h is left-continuous, then

$$\Phi_h(\Pi) = \int_0^1 Q_\Pi(1 - p) \, \mathrm{d}h(p). \tag{1}$$

In what follows, h' represents the right-derivative of h, which exists on $[0, 1)$ since h is concave on $[0, 1]$. Let $\|h'\|_2 = (\int_0^1 (h'(p))^2 \, \mathrm{d}p)^{1/2}$. We give a lemma which we will rely on when considering the EMV problem formulated by Wang and Zhou (2020).

LEMMA 2.2 Theorem 3.1 of Liu *et al.* 2020 *If h is continuous and not constantly zero, then a maximizer Π^* to the optimization problem*

$$\max_{\Pi \in \mathcal{M}^2} \Phi_h(\Pi) \quad \textit{subject to } \mu(\Pi) = m \textit{ and } \sigma^2(\Pi) = s^2 \tag{2}$$

has the following quantile function

$$Q_{\Pi^*}(p) = m + s\frac{h'(1 - p)}{\|h'\|_2}, \quad \textit{a.e. } p \in (0, 1), \tag{3}$$

and the maximum value of (2) *is $\Phi_h(\Pi^*) = s\|h'\|_2$.*

By Lemma 2.2, Han *et al.* (2023) presented many examples linking specific exploratory distributions with the corresponding Choquet regularizers and generated some common exploration measures including ε-greedy, three-point, exponential, uniform and Gaussian; see their Examples 4.1–4.6 and sections 4.3–4.5.

REMARK 2.3 The result in Lemma 2.2 can be extended to a more general case involving higher moments. For $a > 1$, Theorem 5 in Pesenti *et al.* (2025) showed that if the uncertain set is given by

$$\mathcal{M}^a(m, v) = \{\Pi \in \mathcal{M}^a : \mu(\Pi) = m \text{ and } \mathbb{E}[|\Pi - m|^a] \leqslant v^a\},$$

the optimization problem $\max_{\Pi \in \mathcal{M}^a} \Phi_h(\Pi)$, for $p \in (0, 1)$, can be solved by

$$Q_\Pi(p) = m + v\frac{|h'(1 - p) - c_{h,b}|^b}{h'(1 - p) - c_{h,b}}[h]_b^{1-b},$$

$$\text{if } h'(1 - p) - c_{h,b} \neq 0, \text{ and } Q_\Pi(p) = m \text{ otherwise.}$$

† We call Φ_h to be location invariant and scale homogeneous if $\Phi_h(\Pi') = \lambda\Phi_h(\Pi)$ where Π' is the distribution of $\lambda X + c$ for $\lambda > 0$, $c \in \mathbb{R}$ and $X \sim \Pi$.

‡ Π_1 is smaller than Π_2 in *convex order*, denoted by $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$, if $\mathbb{E}[f(\Pi_1)] \leqslant \mathbb{E}[f(\Pi_2)]$ for all convex functions f, provided that the two expectations exist. It is immediate that $\Pi_1 \preceq_{\mathrm{cx}} \Pi_2$ implies $\mathbb{E}[\Pi_1] \leqslant \mathbb{E}[\Pi_2]$.

Here, $b \in [1, \infty]$ is the Hölder conjugate of a, namely $b = (1 - 1/a)^{-1}$, or equivalently, $1/a + 1/b = 1$,

$$c_{h,b} = \arg\min_{x \in \mathbb{R}} \left\| h' - x \right\|_b \quad \text{and}$$
$$[h]_b = \min_{x \in \mathbb{R}} \left\| h' - x \right\|_b = \left\| h' - c_{h,b} \right\|_b,$$

with

$$\left\| h' - x \right\|_b = \left(\int_0^1 \left| h'(p) - x \right|^b \mathrm{d}p \right)^{1/b}, \ b < \infty \quad \text{and}$$
$$\left\| h' - x \right\|_\infty = \max_{p \in [0,1]} \left| h'(p) - x \right|, \ x \in \mathbb{R}.$$

2.2. *Continuous-time EMV problem*

The classical MV problem has been well studied in the literature; see e.g. Markowitz (1952), Li and Ng (2000) and Li *et al.* (2002). We first briefly introduce the classical MV problem in continuous time.

Let T be a fixed investment planning horizon and $\{W_t, 0 \leqslant t \leqslant T\}$ be a standard Brownian motion defined on a given filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{0 \leqslant t \leqslant T}, \mathbb{P})$ that satisfies usual conditions. Assume that a financial market consists of a riskless asset and only one risky asset, where the riskless asset has a constant interest rate $r > 0$ and the risky asset has a price process governed by

$$\mathrm{d}S_t = S_t(\mu \, \mathrm{d}t + \sigma \, \mathrm{d}W_t), \quad 0 \leqslant t \leqslant T, \tag{4}$$

with $S_0 = s_0 > 0$ where $\mu \in \mathbb{R}, \sigma > 0$ is the mean and volatility parameters, respectively. The Sharpe ratio of the risky asset is defined by $\rho = (\mu - r)/\sigma$. Let $u = \{u_t, 0 \leqslant t \leqslant T\}$ denote the discounted amount invested in the risky asset at time t, and the rest of the wealth is invested in the risk-free asset. By (4), the discounted wealth process $\{X_t^u, 0 \leqslant t \leqslant T\}$ for a policy u_t is then given as

$$\mathrm{d}X_t^u = \sigma u_t(\rho \, \mathrm{d}t + \mathrm{d}W_t), \quad 0 \leqslant t \leqslant T, \tag{5}$$

with $X_0^u = x_0 \in \mathbb{R}$. Under the continuous-time MV setting, we aim to solve the following constrained optimization problem

$$\min_u \mathrm{Var}[X_T^u] \quad \text{subject to } \mathrm{E}[X_T^u] = z, \tag{6}$$

where $\{X_t^u, 0 \leqslant t \leqslant T\}$ satisfies the dynamics (5) under the investment policy u, and $z \in \mathbb{R}$ is an investment target determined at $t = 0$ as the desired mean payoff at the end of the investment horizon $[0, T]$.

By applying a Lagrange multiplier w, we can transform (6) into an unconstrained problem

$$\min_u \mathrm{E}[(X_T^u)^2] - z^2 - 2w(\mathrm{E}[X_T^u] - z)$$
$$= \min_u \mathrm{E}[(X_T^u - w)^2] - (w - z)^2. \tag{7}$$

The problem in (7) was well studied by Li and Ng (2000), and it can be solved analytically, whose solution u^* depends on w. Then the original constraint $\mathrm{E}[X_T^{u^*}] = z$ determines the value of w.

Employing the method in Wang *et al.* (2020a) and Wang and Zhou (2020), we give the 'exploratory' version of the state dynamic (5) motivated by repetitive learning in RL. In this formulation, the control process is now randomized, leading to a distributional or exploratory control process denoted by $\Pi = \{\Pi_t, 0 \leqslant t \leqslant T\}$. Here, $\Pi_t \in \mathscr{M}(U)$ is the probability distribution function for control at time t, with $\mathscr{M}(U)$ being the set of distribution functions on U. For such a given distributional control $\Pi \in \mathscr{M}(U)$, the exploratory version of the state dynamics in (5) is changed to

$$\mathrm{d}X_t^\Pi = \widetilde{b}(\Pi_t) \, \mathrm{d}t + \widetilde{\sigma}(\Pi_t) \, \mathrm{d}W_t, \quad 0 < t \leqslant T, \tag{8}$$

with $X_0^\Pi = x_0$, where

$$\widetilde{b}(\Pi) := \int_{\mathbb{R}} \rho\sigma u \, \mathrm{d}\Pi(u) \quad \text{and} \quad \widetilde{\sigma}(\Pi) := \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \, \mathrm{d}\Pi(u)}. \tag{9}$$

Denote the mean and variance processes associated with the control process Π by μ_t and σ_t^2 for $0 \leqslant t \leqslant T$:

$$\mu_t := \int_{\mathbb{R}} u \, \mathrm{d}\Pi_t(u), \quad \sigma_t^2 := \int_{\mathbb{R}} u^2 \, \mathrm{d}\Pi_t(u) - \mu_t^2. \tag{10}$$

Then it follows from (8)–(10) that

$$\mathrm{d}X_t^\Pi = \rho\sigma\mu_t \, \mathrm{d}t + \sigma\sqrt{\mu_t^2 + \sigma_t^2} \, \mathrm{d}W_t, \tag{11}$$

with $X_0^\Pi = x_0$. We refer to Wang *et al.* (2020a, pp. 6–8) for more detailed explanation of where this exploratory formulation comes from.

Next, we use a Choquet regularizer Φ_h to measure the level of exploration, and the aim of the exploratory control is to achieve a continuous-time EMV problem under the framework of RL. For any fixed $w \in \mathbb{R}$, we get the Choquet-regularized EMV problem by adding an exploration weight $\lambda > 0$, which reflects the strength of the exploration desire:

$$\min_{\Pi \in \mathscr{A}(0,x_0)} \mathrm{E}\left[(X_T^\Pi - w)^2 - \lambda \int_0^T \Phi_h(\Pi_t) \, \mathrm{d}t \right] - (w - z)^2,$$

where $\mathscr{A}(t,x)$ is the set of all admissible controls Π for $(t,x) \in [0,T) \times \mathbb{R}$. A control process $\Pi \in \mathscr{A}(t,x)$ is said to be admissible if (i) for $t \leqslant s \leqslant T$, $\Pi_s \in \mathscr{M}(\mathbb{R})$ a.s.; (ii) for $A \in \mathscr{B}(\mathbb{R})$, $\{\int_A \Pi_s(u) \, \mathrm{d}u, t \leqslant s \leqslant T\}$ is $\mathscr{F}_s$-progressively measurable; (iii) $\mathrm{E}[\int_t^T (\mu_s^2 + \sigma_s^2) \, \mathrm{d}s] < \infty$; and (iv) $\mathrm{E}[(X_T^\Pi - w)^2 - \lambda \int_t^T \Phi_h(\Pi_s) \, \mathrm{d}s \mid X_t^\Pi = x] < \infty$.

The value function is then defined as

$$V(t,x;w) := \inf_{\Pi \in \mathscr{A}(t,x)} \mathrm{E}\left[(X_T^\Pi - w)^2 - \lambda \int_t^T \Phi_h(\Pi_s) \, \mathrm{d}s \mid X_t^\Pi = x \right] - (w - z)^2, \tag{12}$$

and the value function under feedback control Π is

$$V^\Pi(t,x;w) := \mathrm{E}\left[(X_T^\Pi - w)^2 - \lambda \int_t^T \Phi_h(\Pi_s) \, \mathrm{d}s \mid X_t^\Pi = x \right] - (w - z)^2. \tag{13}$$

3. Solving EMV problem

In this section, we aim to solve the Choquet-regularized EMV problem. Firstly, we have following result based on Lemma 2.2.

PROPOSITION 3.1 *Let a continuous $h \in \mathcal{H}$ be given. For any $\Pi = \{\Pi_t\}_{t\geqslant 0} \in \mathcal{A}(t,x)$ with mean process $\{\mu_t\}_{t\geqslant 0}$ and variance process $\{\sigma_t^2\}_{t\geqslant 0}$, there exists $\Pi^* = \{\Pi_t^*\}_{t\geqslant 0} \in \mathcal{A}(t,x)$ given by*

$$Q_{\Pi_t^*}(p) = \mu_t + \sigma_t \frac{h'(1-p)}{\|h'\|_2}, \quad a.e.\ p \in (0,1), \quad t \geqslant 0, \tag{14}$$

which has the same mean and variance processes satisfying $V^{\Pi^*}(t,x;w) \leqslant V^{\Pi}(t,x;w)$.

Proof By (8), it is clear that the term $\mathrm{E}[(X_T^{\Pi} - w)^2 | X_t^{\Pi} = x]$ in (13) only depends on the mean process $\{\mu_t\}_{t\geqslant 0}$ and the variance process $\{\sigma_t^2\}_{t\geqslant 0}$ of $\{\Pi_t\}_{t\geqslant 0}$. Thus, for any fixed $t \geqslant 0$, choose Π_t^* with mean μ_t and variance σ_t^2 that maximizes $\Phi_h(\Pi_t)$. Together with Lemma 2.2, we get the desired result. ■

Proposition 3.1 indicates that the control problem in (12) is minimized within a location–scale family of distributions,† which is determined only by h. In fact, if $\hat{\Pi}_t$ is in the location-scale family of Π_t^*, then we have $\hat{\Pi}_t(x) = \Pi_t^*((x-a)/b)$ for some $a \in \mathbb{R}$ and $b > 0$ for all $x \in \mathbb{R}$. Since $\Phi_{\lambda h}(\Pi) = \lambda \Phi_h(\Pi)$ for any $\lambda > 0$, Π that maximizes Φ_h also maximizes $\Phi_{\lambda h}$. Thus, by Proposition 3.1, we have $h'(p) = Q_{\Pi_t^*}(1-p) - \mu_t = (Q_{\hat{\Pi}_t^*}(1-p) - a)/b - \mu_t$ for $p \in (0,1)$ a.e. Since $\mu(\hat{\Pi}_t) = a + b\mu_t$, it follows that $\hat{\Pi}_t$ maximizes Φ_h over $\mathcal{M}^2(\mu_t(\hat{\Pi}), \sigma_t^2(\hat{\Pi}))$.

REMARK 3.2 We know from Remark 2.3 that if both the reward term and the dynamic process only depend on the mean process μ_t and the a-th moment process σ_t^a of Π_t for $t \geqslant 0$, then we have $V^{\Pi^*}(t,x;w) \leqslant V^{\Pi}(t,x;w)$ with Π_t^* satisfying

$$Q_{\Pi_t^*}(p) = \mu_t + \sigma_t \frac{\left|h'(1-p) - c_{h,b}\right|^b}{h'(1-p) - c_{h,b}} [h]_b^{1-b},$$
$$\text{if } h'(1-p) - c_{h,b} \neq 0, \quad \text{and}$$
$$Q_{\Pi_t^*}(p) = \mu_t, \text{ otherwise.}$$

Using the Bellman's dynamic principle, we get

$$V(t,x;w) = \inf_{\Pi \in \mathcal{A}(t,x)} \mathrm{E}\left[-\lambda \int_t^s \Phi_h(\Pi_v)\, \mathrm{d}v + V(s, X_s^{\Pi}; w) \,|\, X_t^{\Pi} = x \right]. \tag{15}$$

Then we can deduce from (15) that V satisfies the HJB equation

$$V_t(t,x;w) + \min_{\Pi \in \mathscr{M}(\mathbb{R})} \left[\frac{1}{2} \widetilde{\sigma}^2(\Pi) V_{xx}(t,x;w) + \widetilde{b}(\Pi) V_x(t,x;w) - \lambda \Phi_h(\Pi) \right] = 0. \tag{16}$$

By (9), the HJB equation in (16) is equivalent to

$$V_t(t,x;w) + \min_{\Pi \in \mathscr{M}(\mathbb{R})} \left[\frac{\sigma^2}{2} \left(\mu(\Pi)^2 + \sigma(\Pi)^2 \right) V_{xx}(t,x;w) + \rho\sigma\mu(\Pi) V_x(t,x;w) - \lambda \Phi_h(\Pi) \right] = 0, \tag{17}$$

with terminal condition $V(T,x;w) = (x-w)^2 - (w-z)^2$. Here, we assume that Π has finite second-order moment, and $\mu(\Pi)$ and $\sigma(\Pi)^2$ are the mean and variance of Π, respectively.

We now pay attention to the minimization in (17). Let

$$\varphi(t,x,\Pi) = \frac{\sigma^2}{2} \left(\mu(\Pi)^2 + \sigma(\Pi)^2 \right) V_{xx}(t,x;w) + \rho\sigma\mu(\Pi) V_x(t,x;w) - \lambda \Phi_h(\Pi).$$

Note that $\varphi(t,x,\Pi)$ only depends on Π by $\mu(\Pi)$ and $\sigma(\Pi)^2$ except $\Phi_h(\Pi)$, we get

$$\min_{\Pi \in \mathscr{M}(\mathbb{R})} \varphi(t,x,\Pi) = \min_{m \in \mathbb{R}, s>0} \min_{\substack{\Pi \in \mathscr{M}(R) \\ \mu(\Pi) = m, \sigma(\Pi)^2 = s^2}} \varphi(t,x,\Pi),$$

and the inner minimization problem is equivalent to

$$\max_{\Pi \in \mathscr{M}(R)} \Phi_h(\Pi) \quad \text{subjectto } \mu(\Pi) = m, \quad \sigma(\Pi)^2 = s^2. \tag{18}$$

By Lemma 2.2, the maximizer Π^* of (18) whose quantile function is $Q_{\Pi^*}(p)$ satisfies

$$Q_{\Pi^*}(p) = m + s \frac{h'(1-p)}{\|h'\|_2}, \tag{19}$$

and $\Phi_h(\Pi^*) = s\|h'\|_2$. Then the HJB equation in (17) is converted to

$$V_t(t,x;w) + \min_{m \in \mathbb{R}, s>0} \left[\frac{\sigma^2}{2} (m^2 + s^2) V_{xx}(t,x;w) + \rho\sigma m V_x(t,x;w) - \lambda s \|h'\|_2 \right] = 0. \tag{20}$$

By the first-order conditions, we get the minimizer of (20)

$$m^* = -\frac{\rho}{\sigma} \frac{V_x}{V_{xx}}, \quad \text{and} \quad s^* = \frac{\lambda \|h'\|_2}{\sigma^2 v_{xx}}. \tag{21}$$

Bringing m^* and s^* back into (20), we can rewrite (20) as

$$V_t - \frac{\rho^2}{2} \frac{V_x^2}{V_{xx}} - \frac{\lambda^2}{2\sigma^2} \frac{\|h'\|_2^2}{V_{xx}} = 0. \tag{22}$$

By the terminal condition $V(T,x;w) = (x-w)^2 - (w-z)^2$, a smooth solution to (22) is given by

$$V(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)}$$

† Recall that given a distribution Π the *location-scale family* of Π is the set of all distributions $\Pi_{a,b}$ parameterized by $a \in \mathbb{R}$ and $b > 0$ such that $\Pi_{a,b}(x) = \Pi((x-a)/b)$ for all $x \in \mathbb{R}$.

$$-\frac{\lambda^2\|h'\|_2^2}{4\rho^2\sigma^2}(e^{\rho^2(T-t)}-1)-(w-z)^2. \tag{23}$$

Then we can deduce from (19), (21) and (23) that

$$m^*=-\frac{\rho}{\sigma}(x-w),\quad \text{and}\quad s^*=\frac{\lambda\|h'\|_2}{2\sigma^2}e^{\rho^2(T-t)},$$

and the dynamic (11) under Π^* becomes

$$\begin{aligned}\mathrm{d}X_t^*&=-\rho^2(X_t^*-w)\,\mathrm{d}t\\&\quad+\sqrt{\rho^2(X_t^*-w)^2+\frac{\lambda^2\|h'\|_2^2}{4\sigma^2}e^{2\rho^2(T-t)}}\,\mathrm{d}W_t\end{aligned}$$

with $X_0^*=x_0$.

Finally, we try to calculate w. By $\mathrm{E}[\max_{t\in[0,T]}(X_t^*)^2]<\infty$ and using Fubini theorem, we get

$$\begin{aligned}\mathrm{E}[X_t^*]&=x_0+\mathrm{E}\left[\int_0^t-\rho^2(X_s^*-w)\,\mathrm{d}s\right]\\&=x_0+\int_0^t-\rho^2(\mathrm{E}[X_s^*]-w)\,\mathrm{d}s.\end{aligned}$$

Hence, $\mathrm{E}[X_t^*]=(x_0-w)^2e^{-\rho^2t}+w$. It follows from $\mathrm{E}[X_T^*]=z$ that

$$w=\frac{ze^{\rho^2T}-x_0}{e^{\rho^2T}-1}.$$

We summarize the above results in the following theorem.

THEOREM 3.3 *The value function of Choquet-regularized EMV problem in* (12) *is given by*

$$\begin{aligned}V(t,x;w)&=(x-w)^2e^{-\rho^2(T-t)}\\&\quad-\frac{\lambda^2\|h'\|_2^2}{4\rho^2\sigma^2}(e^{\rho^2(T-t)}-1)-(w-z)^2,\end{aligned}\tag{24}$$

and the corresponding optimal control process is Π^**, whose quantile function is*

$$Q_{\Pi^*}(p)=-\frac{\rho}{\sigma}(x-w)+\frac{\lambda h'(1-p)}{2\sigma^2}e^{\rho^2(T-t)},\tag{25}$$

with the mean and variance of Π^*

$$\mu(\Pi^*)=-\frac{\rho}{\sigma}(x-w),\quad\text{and}\quad\sigma(\Pi^*)^2=\frac{\lambda^2\|h'\|_2^2}{4\sigma^4}e^{2\rho^2(T-t)}.\tag{26}$$

The optimal wealth process under Π^* *is the unique solution of the SDE*

$$\begin{aligned}\mathrm{d}X_t^*&=-\rho^2(X_t^*-w)\,\mathrm{d}t\\&\quad+\sqrt{\rho^2(X_t^*-w)^2+\frac{\lambda^2\|h'\|_2^2}{4\sigma^2}e^{2\rho^2(T-t)}}\,\mathrm{d}W_t\end{aligned}$$

with $x_0^*=x_0$. *Finally, the Lagrange multiplier* w *is given by*

$$w=\frac{ze^{\rho^2T}-x_0}{e^{\rho^2T}-1}.$$

Proof Along with the similar lines of the verification theorem in Wang *et al.* (2020a) (see their Theorem 4), we can verify that for any $w\in\mathbb{R}$, (24) is indeed the value function and the optimal control Π^* is admissible. ■

There are several observations to note in this result. We can see from (25) that for any Choquet regularizer, the optimal exploratory distribution is uniquely determined by h'. Different h corresponds to a different Choquet regularizer; hence h will certainly affect the way and the level of exploration. Also, since $h'(x)$ is the 'probability weight' put on x when calculating the (nonlinear) Choquet expectation; see e.g. Quiggin (1982) and Gilboa and Schmeidler (1989), the more weight put on the level of exploration, the more spread out the exploration becomes around the current position. In addition, we point out that if we fix the value of $\|h'\|_2^2$ for different Choquet regularizers by multiplying or dividing by a constant, the mean and variance of the different optimal distributions are equal.

Moreover, the optimal control processes under Φ_h has the same expectation as the one in Wang and Zhou (2020) when the differential entropy is used as a regularizer, which is also identical to the optimal control of the classical, non-exploratory MV problem, and the expectation is independent of λ and h. Meanwhile, the variance of optimal control process is independent of state x but decreases over time, which is different from Han *et al.* (2023) where an infinite horizon counterpart is studied. This is intuitive because by exploration, one can get more information over time, and then the demand and aspiration of exploration decreases. In a sense, the expectation represents exploitation which means making the best decision based on existing information, and the variance represents exploration. As a result, the observations above show a perfect separation between exploitation and exploration.

In the following example, we show optimal exploration samplers under the EMV framework for some concrete choices of h studied in Han *et al.* (2023). Theorem 3.3 yields that the mean of the optimal distribution is independent of h, so we will specify only its quantile function and variance for each h discussed below.

EXAMPLE 3.1 (i) Let $h(p)=-p\log(p)$. Then we have

$$\begin{aligned}&\Phi_h(\Pi)\\&\quad=\int_0^\infty\Pi([x,\infty))\log(\Pi([x,\infty)))\,\mathrm{d}x,\end{aligned}$$

which is the cumulative residual entropy defined in Rao *et al.* (2004) and Hu and Chen (2020); see Example 4.5 of Han *et al.* (2023). The optimal policy is a shifted-exponential distribution given as

$$\begin{aligned}&\Pi^*(u;t,x)\\&\quad=1-\exp\left\{-\frac{2\sigma^2}{\lambda e^{\rho^2(T-t)}}\left(u+\frac{\rho}{\sigma}(x-w)\right)-1\right\}.\end{aligned}$$

Since $\|h'\|_2^2=1$, the variance of Π^* is given by

$$(\sigma^*(x))^2=\frac{\lambda^2}{4\sigma^4}e^{2\rho^2(T-t)}.$$

(ii) Let $h(p) = \int_0^p z(1-s)\mathrm{d}s$, where z is the standard normal quantile function. We have $\Phi_h(\Pi) = \int_0^1 Q_\Pi(p)z(p)\,\mathrm{d}p$; see Example 4.6 of Han *et al.* (2023). The optimal policy is a normal distribution given by

$$\Pi^*(\cdot;t,x) = \mathrm{N}\left(-\frac{\rho}{\sigma}(x-w), \frac{\lambda^2}{4\sigma^4}e^{2\rho^2(T-t)}\right),$$

owing to the fact that $\|h'\|_2^2 = 1$.

(iii) Let $h(p) = p - p^2$. Then $\Phi_h(\Pi) = \mathbb{E}[|X_1 - X_2|]/2$, which is the Gini mean difference; see Section 4.5 of Han *et al.* (2023). The optimal policy $\Pi^*(\cdot;x)$ is a uniform distribution given as

$$\mathrm{U}\left[-\frac{\rho}{\sigma}(x-w) - \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}, -\frac{\rho}{\sigma}(x-w) + \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\right].$$

Since $\|h'\|_2^2 = 1/3$, the variance of Π^* is given by $(\sigma^*(x))^2 = \lambda^2 e^{2\rho^2(T-t)}/12\sigma^4$.

REMARK 3.4 For $h \in \mathcal{H}$ with a general Π, we can convert sampling from Π to sampling from a uniform distribution. To be specific, assuming ξ is a uniform random variable on $[0,1]$, we then have $\mathbb{P}(Q_\Pi(\xi) \leqslant a) = \mathbb{P}(\xi \leqslant \Pi(a)) = \Pi(a)$. Consequently, $Q_\Pi(\xi)$ follows the distribution Π.

4. An alternative form of Choquet regularizers

As mentioned in Introduction, for an absolutely continuous Π, Shannon's differential entropy, defined as

$$\mathrm{DE}(\Pi) := -\int_{\mathbb{R}} \Pi'(x)\log(\Pi'(x))\,\mathrm{d}x \tag{27}$$

is commonly used for exploration–exploitation balance in RL; see Wang and Zhou (2020), Guo *et al.* (2022), Jiang *et al.* (2022) and Dai *et al.* (2023). It admits a different quantile representation (see Sunoj and Sankaran 2012)

$$\mathrm{DE}(\Pi) = \int_0^1 \log(Q'_\Pi(p))\,\mathrm{d}p.$$

It is clear that DE is location invariant, but not scale homogeneous. It is not quantile additive either. Therefore, DE is *not* a Choquet regularizer.

Inspired by the logarithmic form of DE, we consider another EMV problem:

$$\widehat{V}(t,x;w) := \inf_{\Pi\in\mathscr{A}(t,x)} \mathrm{E}\left[(X_T^\Pi - w)^2 - \lambda\int_t^T \log\Phi_h(\Pi_s)\,\mathrm{d}s \mid X_t^\Pi = x\right] - (w-z)^2, \tag{28}$$

where we apply the logarithmic form of Φ_h as the regularizer to measure and manage the level of exploration. According to the monotonicity and concavity of logarithmic function, we can easily verify that $\log\Phi_h$ is still a concave mapping:

$$\begin{aligned}&\log\Phi_h(\lambda\Pi_1 + (1-\lambda)\Pi_2)\\ &\quad\geqslant \log(\lambda\Phi_h(\Pi_1) + (1-\lambda)\Phi_h(\Pi_2))\\ &\quad\geqslant \lambda\log\Phi_h(\Pi_1) + (1-\lambda)\log\Phi_h(\Pi_2)\end{aligned}$$

for all $\Pi_1, \Pi_2 \in \mathcal{M}$ and $\lambda \in [0,1]$, and consistent with convex order:

$$\log\Phi_h(\Pi_1) \leqslant \log\Phi_h(\Pi_2), \quad \text{for all } \Pi_1, \Pi_2 \in \mathcal{M} \text{ with } \Pi_1 \preceq_{\mathrm{cx}} \Pi_2.$$

Comparing to the properties of Φ_h, $\log\Phi_h$ is not necessarily non-negative as Φ_h. However, the non-negativity does not inherently affect the exploration. Further, $\Phi(\Pi)$ is zero when Π is Dirac measure, we then have $\log\Phi(\delta_c) = -\infty$ for all $c \in \mathbb{R}$. The location invariance for $\log\Phi_h$ is obvious. For scale homogeneity, $\log\Phi_h$ is no longer linear in its scale, but we have $\log\Phi_h(\Pi') = \log\Phi_h(\Pi) + \log\lambda$ for any $\lambda > 0$ where Π' is the distribution of λX for $\lambda > 0$ and $X \sim \Pi$. It is interesting to see that the level of randomness is captured by the term of $\log\lambda$. Based on the observations above, we find that $\log\Phi_h$ has many similarities with DE in capturing the randomness.

We remark that maximizing Φ_h over $\mathcal{M}^2(m,s^2)$ is equivalent to maximizing $\log\Phi_h$ over $\mathcal{M}^2(m,s^2)$. In the following theorem, we give the optimal result of (28) directly. Since the procedure is similar to section 3, we omit the details here.

THEOREM 4.1 *The value function of* (28) *is given by*

$$\widehat{V}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2-t^2) - \frac{\lambda}{2}\left(\rho^2 T + \log\frac{\lambda\|h'\|_2^2}{2e\sigma^2}\right)(T-t) - (w-z)^2, \tag{29}$$

and the corresponding optimal control process is $\widehat{\Pi}^*$ *with quantile function*

$$Q_{\widehat{\Pi}^*}(p) = -\frac{\rho}{\sigma}(x-w) + \sqrt{\frac{\lambda}{2\sigma^2\|h'\|_2^2}}e^{\frac{1}{2}\rho^2(T-t)}h'(1-p). \tag{30}$$

Moreover, the mean and variance of Π^* *are*

$$\mu(\widehat{\Pi}^*) = -\frac{\rho}{\sigma}(x-w), \quad \text{and} \quad \sigma(\widehat{\Pi}^*)^2 = \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}.$$

The optimal wealth process under Π^* *is the unique solution of the SDE*

$$\mathrm{d}X_t^* = -\rho^2(X_t^* - w)\,\mathrm{d}t + \sqrt{\rho^2(X_t^*-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}\,\mathrm{d}W_t$$

with $X_0^* = x_0$. *Finally, the Lagrange multiplier* w *is given by*

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}.$$

REMARK 4.2 By (30), we can see that the optimal exploratory distribution is also uniquely determined by h'. Since the form of $\log \Phi_h$ affects the value function, even though the form of optimal distributions is the same, it is to be expected that the variance of the optimal distributions is different from (25). It is worth pointing that the mean and variance of the optimal distributions are the same as those in Wang and Zhou (2020) where the differential entropy is used as a regularizer, which is an interesting observation. This is because for the payoff function depending only on the mean and variance processes of the distributional control, the Gaussian distribution maximizes the entropy when the mean and variance are fixed, and the maximized MV constrained entropy and $\log \Phi_h$ are equal up to a constant and both logarithmic in the given standard deviation and independent of the mean. Moreover, since different h corresponds to different exploratory distributions, our optimal exploratory distributions are no longer necessarily Gaussian as in Wang and Zhou (2020), and are dictated by the choice of Choquet regularizers, which can be such as Gaussian, uniform distribution or exponential distribution.

Parallel to Example 3.1, we give Example 4.1. Theorem 4.1 yields that both the mean and the variance of the optimal distribution are independent of h, so we will specify only its quantile function.

EXAMPLE 4.1 (i) Let $h(p) = -p\log(p)$. Then we have

$$\log \Phi_h(\Pi) = \log \int_0^\infty \Pi([x,\infty)) \log(\Pi([x,\infty))) \, \mathrm{d}x.$$

The optimal policy is a shifted-exponential distribution given as

$$\Pi^*(u;t,x) = 1 - \exp\left\{ -\sqrt{\frac{2\sigma^2}{\lambda e^{\rho^2(T-t)}}}\left(u + \frac{\rho}{\sigma}(x-w)\right) - 1 \right\}.$$

(ii) Let $h(p) = \int_0^p z(1-s)\,\mathrm{d}s$, where z is the standard normal quantile function. We have $\log \Phi_h(\Pi) = \log \int_0^1 Q_\Pi(p) z(p)\,\mathrm{d}p$. The optimal policy is a normal distribution given by

$$\Pi^*(\cdot;t,x) = \mathrm{N}\left(-\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)} \right).$$

(iii) Let $h(p) = p - p^2$. Then $\log \Phi_h(\Pi) = \log \mathbb{E}[|X_1 - X_2|] - \log 2$. The optimal policy $\Pi^*(\cdot;x)$ is a uniform distribution given as

$$\mathrm{U}\left[-\frac{\rho}{\sigma}(x-w) - \sqrt{\frac{3\lambda}{2\sigma^2} e^{\rho^2(T-t)}}, -\frac{\rho}{\sigma}(x-w) + \sqrt{\frac{3\lambda}{2\sigma^2} e^{\rho^2(T-t)}} \right].$$

REMARK 4.3 The optimal policy in Example 4.1(ii) coincides with the entropy-regularized optimal policy derived in Wang and Zhou (2020), where the authors established the optimality of Gaussian exploration distributions under Shannon entropy regularization. Specifically, for a Gaussian distribution $\Pi = \mathrm{N}(m, s^2)$, the Shannon entropy regularization term defined in (27) takes the form $\mathrm{DE}(\Pi) = \log s + \frac{1}{2}\log(2\pi e)$. We can also compute the logarithmic Choquet integral with the distortion function $h_1(p) = \int_0^p z(1-s)\,\mathrm{d}s$, yielding $\log \Phi_{h_1}(\Pi) = \log s$. This reveals that the original optimization problem under entropy regularization is equivalent to the logarithmic Choquet-regularized counterpart with distortion h_1, showing that $\log \Phi_h$ generalizes Shannon entropy regularization as a special case. Therefore, the logarithmic Choquet integral offers greater flexibility by accommodating a broader class of distortion functions h.

Next, we consider the solvability equivalence between the classical and the exploratory MV problems. Here, 'solvability equivalence' implies that the solution of one problem will lead to that of the other directly, without needing to solve it separately. Recall the classical MV problem in section 2.2. The explicit forms of optimal control and value function, denoted respectively by u^* and V^{cl}, were given by Theorem 3.2-(b) of Wang and Zhou (2020). We provide the solvability equivalence between the classical and the exploratory MV problems defined by (7), (12) and (28), respectively. Since the proof is similar to that of Theorem 9 in Appendix C of Wang *et al.* (2020a), we omit the details here.

PROPOSITION 4.4 *The following three statements (a), (b), (c) are equivalent.*

(a) *The function* $V(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - \frac{\lambda^2 \|h'\|_2^2}{4\rho^2\sigma^2}(e^{\rho^2(T-t)} - 1) - (w-z)^2$, $(t,x) \in [0,T]\times\mathbb{R}$, *is the value function of the EMV problem* (12) *and the optimal feedback control is* Π^*, *whose quantile function is*

$$Q_{\Pi^*}(p) = -\frac{\rho}{\sigma}(x-w) + \frac{\lambda h'(1-p)}{2\sigma^2} e^{\rho^2(T-t)}.$$

(b) *The value function* $\widehat{V}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\left(\rho^2 T + \log\frac{\lambda\|h'\|_2^2}{2e\sigma^2}\right)(T-t) - (w-z)^2$, $(t,x) \in [0,T]\times\mathbb{R}$, *is the value function of the EMV problem (28) and the optimal feedback control is* $\widehat{\Pi}^*$, *whose quantile function is*

$$Q_{\widehat{\Pi}^*}(p) = -\frac{\rho}{\sigma}(x-w) + \sqrt{\frac{\lambda}{2\sigma^2\|h'\|_2^2}} h'(1-p) e^{\frac{1}{2}\rho^2(T-t)}.$$

(c) *The function* $V^{cl}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2$, $(t,x) \in [0,T]\times\mathbb{R}$, *is the value function of the classical MV problem* (7) *and the optimal feedback control is*

$$u^*(t,x;w) = -\frac{\rho}{\sigma}(x-w).$$

Moreover, the three problems above all have the same Lagrange multiplier

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}.$$

From the proposition above, we naturally want to explore more connections between (a), (b) and (c). In fact, they have the following convergence property.

PROPOSITION 4.5 *Suppose that statement (a) or (b) or (c) of Proposition* 4.4 *holds. Then for each* $(t, x, w) \in [0, T] \times \mathbb{R} \times \mathbb{R}$,

$$\lim_{\lambda \to 0} \widehat{\Pi}^*(\cdot; t, x; w) = \lim_{\lambda \to 0} \Pi^*(\cdot; t, x; w) = \delta_{u^*(t,x;w)}(\cdot) \text{ weakly},$$

and

$$\lim_{\lambda \to 0} |V(t, x; w) - V^{cl}(t, x; w)| = 0,$$
$$\text{and} \quad \lim_{\lambda \to 0} |\widehat{V}(t, x; w) - V^{cl}(t, x; w)| = 0.$$

Proof The weak convergence is obvious and the convergence of value function follows from

$$\lim_{\lambda \to 0} \frac{\lambda^2 \|h'\|_2^2}{4\rho^2 \sigma^2} (e^{\rho^2 (T-t)} - 1) = 0,$$
$$\text{and} \quad \lim_{\lambda \to 0} \frac{\lambda}{2} \log \frac{\lambda \|h'\|_2^2}{2e\sigma^2} = 0.$$

■

Next, we examine the 'cost of exploration' – the loss in the original (i.e. non-regularized) objective due to exploration, which was originally defined and derived in Wang *et al.* (2020a) for problems with entropy regularization. Due to the explicit inclusion of exploration in the objectives (12) and (28), the cost of the EMV problems are defined as

$$C^{u^*,\Pi^*}(0, x_0; w) = \left(V(0, x_0; w) + \lambda \mathbb{E}\left[\int_0^T \Phi_h(\Pi_t^*) dt | X_0^{\Pi^*} = x_0 \right] \right) - V^{cl}(0, x_0; w), \tag{31}$$

and

$$\widehat{C}^{u^*,\widehat{\Pi}^*}(0, x_0; w) = \left(\widehat{V}(0, x_0; w) + \lambda \mathbb{E}\left[\int_0^T \log \Phi_h(\widehat{\Pi}_t^*) \, \mathrm{d}t \, | \, X_0^{\widehat{\Pi}^*} = x_0 \right] \right) - V^{cl}(0, x_0; w). \tag{32}$$

PROPOSITION 4.6 *Suppose that statement (a) or (b) or (c) of Proposition* 4.4 *holds. Then the cost of exploration for the EMV problem are, respectively, given as*

$$C^{u^*,\Pi^*}(0, x_0; w) = \frac{\lambda^2 \|h'\|_2^2}{4\rho^2 \sigma^2} (e^{\rho^2 T} - 1), \tag{33}$$

and

$$\widehat{C}^{u^*,\widehat{\Pi}^*}(0, x_0; w) = \frac{\lambda T}{2}. \tag{34}$$

Proof Note that

$$\Phi_h(\Pi_t^*) = \sigma(\Pi_t^*) \|h'\|_2 = \frac{\lambda \|h'\|_2^2}{2\sigma^2} e^{\rho^2 (T-t)},$$

and

$$\log \Phi_h(\widehat{\Pi}_t^*) = \log \left(\sigma(\widehat{\Pi}_t^*) \|h'\|_2 \right) = \frac{1}{2} \log \left(\frac{\lambda \|h'\|_2^2}{2\sigma^2} e^{\rho^2 (T-t)} \right).$$

Bringing $\Phi_h(\Pi_t^*)$ and $\log \Phi_h(\widehat{\Pi}_t^*)$ back into (31) and (32), respectively, we can get (33) and (34). ■

REMARK 4.7 The costs of exploration for the two EMV problems are quite different. When Φ_h is regarded as the regularizer, the derived exploration cost does depend on the unknown model parameters through h, μ and σ. (33) implies that, with other parameters being equal, to reduce the exploration cost one should choose regularizers with smaller values of $\|h'\|_2$. Moreover, by (26), we have

$$C^{u^*,\Pi^*}(0, x_0; w) = \frac{\lambda \|h'\|_2}{2\rho^2} \sigma^*(x_0) - \frac{\lambda^2 \|h'\|_2^2}{4\rho^2 \sigma^2},$$

meaning that the cost is proportional to the standardized deviation of the exploratory control, but inversely proportional to the square of the Sharp ratio ρ^2. In contrast, when $\log \Phi_h$ is regarded as the regularizer, the derived exploration cost only depends on λ and T. It is also interesting to note that $\widehat{C}^{u^*,\Pi^*}(0, x_0; w)$ in (34) is the same as the one using DE as the regularizer; see Theorem 3.4 of Wang and Zhou (2020).

Nevertheless, they also have some common features. The exploration cost increases as the exploration weight λ and the exploration horizon T increase, due to more emphasis placed on exploration. In addition, the costs are both independent of the Lagrange multiplier, which suggests that the exploration cost will not increase when the agent is more aggressive (or risk-seeking) reflected by the expected target z or equivalently the Lagrange multiplier w.

REMARK 4.8 To compare $C^{u^*,\Pi^*}(0, x_0; w)$ and $\widehat{C}^{u^*,\Pi^*}(0, x_0; w)$, we have

$$\frac{C^{u^*,\Pi^*}(0, x_0; w)}{\widehat{C}^{u^*,\widehat{\Pi}^*}(0, x_0; w)} = \frac{\lambda \|h'\|_2^2}{2\sigma^2} \frac{e^{\rho^2 T} - 1}{\rho^2 T} = \frac{\lambda \|h'\|_2^2}{2\sigma^2} \left(1 + \sum_{n=1}^{\infty} \frac{\rho^{2n} T^n}{(n+1)!} \right).$$

Then we can easily verify which regularizer has smaller exploration cost under determined market parameters. In general, from a cost point of view, when λ, $\|h'\|_2$ and ρ^2 are small enough and σ is relatively large, Φ_h is a good choice to reduce cost; otherwise $\log \Phi_h$ may be a better choice.

5. RL algorithm design

5.1. Policy improvement

In RL setting, the policy improvement is an important process which ensures the existence of a new policy better than

any given policy. In Proposition 3.1, we have showed that the EMV problem in (12) can be minimized within a location-scale family of distributions. Such a property is also applied to the EMV problem in (28) when $\log \Phi_h$ is regarded as the regularizer. In the following theorem, by Itô's formula, we can also verify that for any given policy, when the regularizer is Φ_h or $\log \Phi_h$, there always exists a better policy in a location-scale family which depends on h. So we can search the optimal exploration distribution only in this location-scale family.

THEOREM 5.1 *Let $w \in \mathbb{R}$ be fixed and Π (resp.$\widehat{\Pi}$) be an arbitrarily given admissible feedback control whose corresponding value function is $V^{\Pi}(t,x;w)$(resp.$\widehat{V}^{\Pi}(t,x;w)$) under regularizer Φ_h(resp.$\log \Phi_h$). Suppose that $V^{\Pi}(t,x;w)$(resp.$\widehat{V}^{\Pi}(t,x;w)$) $\in C^{1,2}([0,T)\times\mathbb{R}\cap C^0([0,T]\times\mathbb{R}))$ and $V^{\Pi}_{xx}(t,x;w)$ (resp.$\widehat{V}^{\Pi}_{xx}(t,x;w)$) >0 for any $(t,x)\in[0,T)\times\mathbb{R}$. Suppose further that the feedback control $\widetilde{\Pi}$(resp.$\widetilde{\widehat{\Pi}}$) whose quantile function is*

$$Q_{\widetilde{\Pi}}(p) = -\frac{\rho}{\sigma}\frac{V^{\Pi}_x}{V^{\Pi}_{xx}} + \frac{\lambda}{\sigma^2 V^{\Pi}_{xx}} h'(1-p) \tag{35}$$

$$\textit{resp.}\quad Q_{\widetilde{\widehat{\Pi}}}(p) = -\frac{\rho}{\sigma}\frac{\widehat{V}^{\widehat{\Pi}}_x}{\widehat{V}^{\widehat{\Pi}}_{xx}} + \sqrt{\frac{\lambda}{\sigma^2\|h'\|_2^2 \widehat{V}^{\widehat{\Pi}}_{xx}}}\, h'(1-p) \tag{36}$$

is admissible. Then

$$V^{\widetilde{\Pi}}(t,x;w) \leqslant V^{\Pi}(t,x;w), \quad (t,x)\in[0,T)\times\mathbb{R},$$

$$\textit{resp.}\quad \widehat{V}^{\widetilde{\widehat{\Pi}}}(t,x;w) \leqslant \widehat{V}^{\widehat{\Pi}}(t,x;w), \quad (t,x)\in[0,T)\times\mathbb{R}.$$

Proof Let $\widetilde{\Pi} = \{\widetilde{\Pi}_s, s\in[t,T]\}$ and $\widetilde{\widehat{\Pi}} = \{\widetilde{\widehat{\Pi}}_s, s\in[t,T]\}$ be the open-loop control generated by the given feedback control policies $\widetilde{\Pi}$ and $\widetilde{\widehat{\Pi}}$, respectively. By assumption, $\widetilde{\Pi}$ and $\widetilde{\widehat{\Pi}}$ are admissible. Applying Itô's formula, we have for any $(t,x)\in[0,T]\times\mathbb{R}$,

$$\begin{aligned}
V^{\Pi}(s,X_s^{\widetilde{\Pi}}) &= V^{\Pi}(t,x) + \int_t^s V_t^{\Pi}(v,X_v^{\widetilde{\Pi}})\,\mathrm{d}v \\
&\quad + \int_t^s V_x^{\Pi}(v,X_v^{\widetilde{\Pi}})\,\mathrm{d}X_v^{\widetilde{\Pi}} \\
&\quad + \frac{1}{2}\int_t^s V_{xx}^{\Pi}(v,X_v^{\widetilde{\Pi}})\,\mathrm{d}<X^{\widetilde{\Pi}},X^{\widetilde{\Pi}}>_v \\
&= V^{\Pi}(t,x) + \int_t^s V_x^{\Pi}(v,X_v^{\widetilde{\Pi}}) \\
&\quad \sigma\sqrt{\mu(\widetilde{\Pi}_v)^2+\sigma(\widetilde{\Pi}_v)^2}\,\mathrm{d}W_v \\
&\quad + \int_t^s \Big[V_t^{\Pi}(v,X_v^{\widetilde{\Pi}}) + \rho\sigma\mu(\widetilde{\Pi}_v)V_x^{\Pi}(v,X_v^{\widetilde{\Pi}}) \\
&\quad + \frac{\sigma^2}{2}(\mu(\widetilde{\Pi}_v)^2+\sigma(\widetilde{\Pi}_v)^2)V_{xx}^{\Pi}(v,X_v^{\widetilde{\Pi}})\Big]\,\mathrm{d}v.
\end{aligned} \tag{37}$$

Let $\tau_n := \inf\{s\geqslant t : \int_t^s \sigma^2 V_x^{\Pi}(v,X_v^{\widetilde{\Pi}})^2(\mu(\widetilde{\Pi}_v)^2+\sigma(\widetilde{\Pi}_v)^2)\,\mathrm{d}v \geqslant n\}$ be a family of stopping times, then substituting $s\wedge\tau_n$ into (37) and taking expectation we get

$$\begin{aligned}
V^{\Pi}(t,x) = \mathrm{E}\Big[& V^{\Pi}(s\wedge\tau_n, X_{s\wedge\tau_n}^{\widetilde{\Pi}}) - \int_t^{s\wedge\tau_n} [V_t^{\Pi}(v,X_v^{\widetilde{\Pi}}) \\
& + \rho\sigma\mu(\widetilde{\Pi}_v)V_x^{\Pi}(v,X_v^{\widetilde{\Pi}}) + \frac{\sigma^2}{2}(\mu(\widetilde{\Pi}_v)^2 \\
& + \sigma(\widetilde{\Pi}_v)^2)V_{xx}^{\Pi}(v,X_v^{\widetilde{\Pi}})]\,\mathrm{d}v \,|\, X_t^{\widetilde{\Pi}} = x\Big].
\end{aligned} \tag{38}$$

On the other hand, by standard argument we have

$$\begin{aligned}
V_t^{\Pi}(t,x) + \rho\sigma\mu(\Pi)V_x^{\Pi}(t,x) + \frac{\sigma^2}{2}(\mu(\Pi)^2 \\
+ \sigma(\Pi)^2)V_{xx}^{\Pi}(t,x) - \lambda\Phi_h(\Pi) = 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
V_t^{\Pi}(t,x) + \min_{\Pi'\in\mathscr{P}(\mathbb{R})}\Big[& \rho\sigma\mu(\Pi')V_x^{\Pi}(t,x) + \frac{\sigma^2}{2}(\mu(\Pi')^2 \\
& + \sigma(\Pi')^2)V_{xx}^{\Pi}(t,x) - \lambda\Phi_h(\Pi')\Big] \leqslant 0.
\end{aligned} \tag{39}$$

By (21), we know $\widetilde{\Pi}$ is the minimizer of (39). Substituting $\widetilde{\Pi}$ into (39) and bringing back to (38) we have

$$\begin{aligned}
V^{\Pi}(t,x) \geqslant \mathrm{E}\Big[& V^{\Pi}(s\wedge\tau_n, X_{s\wedge\tau_n}^{\widetilde{\Pi}}) \\
& - \int_t^{s\wedge\tau_n} \lambda\Phi_h(\widetilde{\Pi}_v)\,\mathrm{d}v \,|\, X_t^{\widetilde{\Pi}} = x\Big].
\end{aligned} \tag{40}$$

Taking $s=T$ in (40) and sending n to ∞, we obtain

$$\begin{aligned}
V^{\Pi}(t,x) &\geqslant \mathrm{E}\Big[V^{\widetilde{\Pi}}(T,X_T^{\widetilde{\Pi}}) - \lambda\int_t^T \Phi_h(\widetilde{\Pi}_v)\,\mathrm{d}v \,|\, X_t^{\widetilde{\Pi}} = x\Big] \\
&= V^{\widetilde{\Pi}}(t,x).
\end{aligned}$$

The proof of regularizer $\log\Phi_h$ is almost the same, so we omit it. ■

THEOREM 5.2 *Let $\Pi^0(u;t,x,w)$ be a feedback control which has quantile function*

$$Q_{\Pi^0}(p) = Q_{\widehat{\Pi}^0}(p) = a(x-w) + c_1 e^{c_2(T-t)}h'(1-p), \tag{41}$$

and $\{\Pi^n(u;t,x,w)\}$ and $\{\widehat{\Pi}^n(u;t,x,w)\}$ be the sequence of feedback controls updated by (35) *and* (36)*, respectively. Denoted by $\{V^{\Pi^n}(t,x;w)\}$ and $\{\widehat{V}^{\widehat{\Pi}^n}(t,x;w)\}$ the sequence of corresponding value functions. Then*

$$\lim_{n\to\infty}\Pi^n(\cdot;t,x,w) = \Pi^*(\cdot;t,x,w) \textit{ weakly},$$

$$\textit{resp.}\quad \lim_{n\to\infty}\widehat{\Pi}^n(\cdot;t,x,w) = \widehat{\Pi}^*(\cdot;t,x,w) \textit{ weakly},$$

and

$$\lim_{n\to\infty} V^{\Pi^n}(t,x;w) = V(t,x;w), \quad (t,x)\in[0,T),$$

$$\textit{resp.}\quad \lim_{n\to\infty}\widehat{V}^{\widehat{\Pi}^n}(t,x;w) = \widehat{V}(t,x;w), \quad (t,x)\in[0,T)\times\mathbb{R},$$

for any $(t,x,w) \in [0,T] \times \mathbb{R} \times \mathbb{R}$, *where* Π^* *and* $\widehat{\Pi}^*$ *in* (25) *and* (30) *are the optimal controls, and* V *and* $\widehat{V}$ *are the value functions given by* (24) *and* (29).

Proof Here we only provide the detailed proof for the case of Φ_h, and the results of $\log \Phi_h$ can be derived in the same way. Let $\{\Pi_s^0\}$ be the open-loop control generated by Π^0. We can verify that $\{\Pi_s^0\}$ is admissible. The dynamic of wealth under Π^0 is

$$\begin{aligned} \mathrm{d}X_t^{\Pi^0} &= \rho\sigma\mu(\Pi^0)\,\mathrm{d}t + \sigma\sqrt{\mu(\Pi^0)^2 + \sigma(\Pi^0)^2}\,\mathrm{d}W_t, \\ X_t^{\Pi^0} &= x, \end{aligned}$$

and the value function under Π^0 is

$$\begin{aligned} V^{\Pi^0}(t,x) = \mathrm{E}\left[\int_t^T -\lambda\Phi_h(\Pi_v^0)\,\mathrm{d}v + (X_T^{\Pi^0} - w)^2 \,|\, X_t^{\Pi^0} = x\right] \\ - (w-z)^2. \end{aligned}$$

By Feynman-Kac formula, we deduce that V^{Π^0} satisfies the following PDE

$$\begin{aligned} V_t(t,x) + \rho\sigma\mu(\Pi^0)V_x(t,x) + \frac{1}{2}\sigma^2(\mu(\Pi^0)^2 \\ + \sigma(\Pi^0)^2)V_{xx}(t,x) - \lambda\Phi_h(\Pi^0) = 0, \end{aligned}$$

with terminal condition $V^{\Pi^0}(T,x) = (x-w)^2 - (w-z)^2$. Solving this equation we obtain

$$V^{\Pi^0}(t,x;w) = (x-w)^2 e^{(2\rho\sigma a + \sigma^2 a^2)(T-t)} + F_0(t),$$

where $F_0(t)$ is a smooth function which only depends on t. Obviously, $V^{\Pi^0}(t,x;w)$ satisfies the conditions of Theorem 5.1, so we can use (35) to obtain Π^1 whose quantile function is

$$Q_{\Pi^1}(p) = -\frac{\rho}{\sigma}(x-w) + \frac{\lambda h'(1-p)}{2\sigma^2 e^{(2\rho\sigma a + \sigma^2 a^2)(T-t)}},$$

with

$$\begin{aligned} \mu(\Pi^1) &= -\frac{\rho}{\sigma}(x-w), \\ \text{and} \quad \sigma^2(\Pi^1) &= \frac{\lambda^2 \|h'\|_2^2}{4\sigma^2 e^{2(2\rho\sigma a + \sigma^2 a^2)(T-t)}}. \end{aligned}$$

By repeating the above program with Π^1, we have

$$V^{\Pi^1}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + F_1(t),$$

where $F_1(t)$ is a smooth function which only depends on t. Using Theorem 5.1 again we obtain Π^2 whose quantile function is

$$Q_{\Pi^2}(p) = -\frac{\rho}{\sigma}(x-w) + \frac{\lambda h'(1-p)}{2\sigma^2} e^{\rho^2(T-t)},$$

with

$$\mu(\Pi^2) = -\frac{\rho}{\sigma}(x-w), \quad \text{and} \quad \sigma^2(\Pi^2) = \frac{\lambda^2\|h'\|_2^2}{4\sigma^4} e^{2\rho^2(T-t)}.$$

By (25)–(26), we know that Π^2 is optimal. ■

Theorem 5.2 shows that, starting from a given initial quantile function, the sequence of feedback controls updated iteratively converges weakly to the optimal control, and the corresponding value functions also converge. This result not only justifies the effectiveness of the method but also implies that, from a learning perspective, such an iterative scheme can be seen as a stable and convergent policy update process. In particular, choosing an initial distribution of the form (41) guarantees convergence of the learning algorithm.

5.2. *The EMV algorithm*

In this section, we aim to solve (12) and (28) by assuming that there is no knowledge about the underlying parameters. One method to overcome this problem is to replace the parameters by their estimations. However, as mentioned in Introduction, the estimations are usually very sensitive to the sample. We will give an offline RL algorithm with a given training data set based on the Actor-Critic algorithm in Konda and Tsitsiklis (2000), Sutton and Barto (2018) and Jia and Zhou (2022b). The Actor-Critic algorithm is essentially a policy-based algorithm, but additionally learns the value function in order to help the policy function learn better. Meanwhile, we use a self-correcting scheme in Wang and Zhou (2020) to learn the Lagrange multiplier w.

Here, we only present the RL algorithm for the case of Φ_h to solve (12). When using $\log \Phi_h$ as the regularizer, we only need to replace Φ_h by $\log \Phi_h$ and modify the parameterization appropriately.

In continuous-time setting, we first discretize $[0,T]$ into N small intervals $[t_i, t_{i+1}]$, $(i = 0,1,\ldots,N-1)$ whose length is equal to $T/N = \Delta t$. We use policy gradient principle to update Actor; and for Critic, Jia and Zhou (2022a) showed that the time-discretized algorithm converges as $\Delta t \to 0$ as long as the corresponding discrete-time algorithms converges, thus we adopt a learning approach of temporal difference error (the TD error; see Doya 2000, Jia and Zhou 2022a). Assume that Π is a given admissible feedback policy and let $\mathscr{D} = \{(t_i, x_{t_i}), i = 0,1,\ldots,N\}$ be a set of samples, the initial sample is $(0, x_0)$, then for $i = 1,2,\ldots,N$, we sample $u_{t_{i-1}}$ from $\Pi_{t_{i-1}}$ and get x_{t_i} at t_i. On the one hand, we have

$$\begin{aligned} V^{\Pi}(t,x) = \mathrm{E}\left[(X_T^{\Pi} - w)^2 - \lambda\int_t^T \Phi_h(\Pi_s)\,\mathrm{d}s \,|\, X_t^{\Pi} = x\right] \\ - (w-z)^2, \end{aligned}$$

so the TD error at t_i is

$$\begin{aligned} \delta_i = -\lambda\Phi_h(\Pi_{t_i})\Delta t + V^{\Pi}(t_{i+1}, X_{t_{i+1}}) - V^{\Pi}(t_i, X_{t_i}), \\ i = 0,1,\ldots,N-1. \end{aligned}$$

On the other hand, based on (24), we can parameterize the Critic value by

$$V^{\theta}(t,x) = (x-w)^2 e^{-\theta_2(T-t)} - \theta_1(e^{\theta_0(T-t)} - 1) - (w-z)^2.$$

This parameterization is well aligned with the inherently linear-quadratic structure of the mean-variance objective,

making it both natural and analytically tractable. We further assess the robustness of this approach empirically in section 6.2. For a single point t_i, we define the loss function as

$$L(\theta) = \frac{1}{2}(U_{t_i} - V^\theta(t_i, X_{t_i}))^2, \tag{42}$$

where U_{t_i} is the estimation of $V(t_i, X_{t_i})$. We take U_{t_i} a bootstrapping estimate $-\lambda\Phi_h(\Pi_{t_i})\Delta t + V^\theta(t_{i+1}, X_{t_{i+1}})$ in (42) as the temporal difference target which will not generate gradient to update the value function automatically. So the gradient of the loss function is

$$\begin{aligned}\nabla_\theta L(\theta) = -(&-\lambda\Phi_h(\Pi_{t_i})\Delta t + V^\theta(t_{i+1}, X_{t_{i+1}}) \\ &- V^\theta(t_i, X_{t_i}))\nabla_\theta V^\theta(t_i, X_{t_i}).\end{aligned} \tag{43}$$

Let α_θ be the learning rate of θ, then by (43), we can get the gradient and the update rule of θ with a set of sample $\mathscr{D}$:

$$\begin{aligned}\nabla\theta = -\sum_{i=0}^{N-1} \frac{\partial V^\theta}{\partial\theta}(t_i, x_{t_i})[&V^\theta(t_{i+1}, x_{t_{i+1}}) \\ &- V^\theta(t_i, x_{t_i}) - \lambda\Phi_h(\Pi_{t_i}^\phi)\Delta t],\end{aligned} \tag{44}$$

and

$$\theta \longleftarrow \theta - \alpha_\theta\nabla\theta. \tag{45}$$

Based on Theorem 5.2, we can parameterize the policy by Π^ϕ with quantile function

$$Q_{\Pi_t^\phi}(p) = -\phi_0(x - w) + e^{\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2(T-t)}h'(1-p).$$

By Lemma 2.3 of Han *et al.* (2023), we know that

$$\Phi_h(\Pi_t^\phi) = \int_0^1 (-\phi_0(x - w) + e^{\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2(T-t)}h'(p)^2)\,\mathrm{d}p.$$

Let $g(t, x; \phi) = \nabla_\phi V^{\Pi^\phi}(t, x)$ be the policy gradient of Π^ϕ and $p(t, \phi) = \Phi_h(\Pi_t^\phi)$, together with Theorem 5 of Jia and Zhou (2022b), $g(t, x; \phi)$ has the following representation:

$$\begin{aligned}&g(t, x; \phi) \\ &= \mathrm{E}\left[\int_t^T \left\{\frac{\partial}{\partial\phi}\log\dot{\Pi}_t^\phi\left(\mathrm{d}V^{\Pi_\phi}(s, X_s^{\Pi^\phi}) - \lambda p(s, \phi)\,\mathrm{d}s\right)\right.\right. \\ &\quad \left.\left. - \lambda\frac{\partial p}{\partial\phi}(s, \phi)\,\mathrm{d}s\right\}\middle| X_t^{\Pi^\phi} = x\right],\end{aligned} \tag{46}$$

where $\dot{\Pi}_t^\phi$ is the density function of Π_t^ϕ. It should be emphasized that Theorem 5 in Jia and Zhou (2022b) is valid only when the support of the exploration distribution is independent of ϕ. In cases where this condition does not hold, additional terms must be incorporated to capture the effect of ϕ on the support set. Let α_ϕ be the learning rate of ϕ, then by (46), we can also get the gradient and the update rule of ϕ with a set of sample $\mathscr{D}$:

$$\begin{aligned}\nabla\phi = \sum_{i=0}^{N-1}\left\{\frac{\partial}{\partial\phi}\log\dot{\Pi}^\phi(u_{t_i}|t_i, x_{t_i})[V^\theta(t_{i+1}, x_{t_{i+1}}) - V^\theta(t_i, x_{t_i})\right. \\ \left. - \lambda\Phi_h(\Pi_{t_i}^\phi)\Delta t] - \lambda\frac{\partial p}{\partial\phi}(t_i, x_{t_i}, \phi)\Delta t\right\},\end{aligned} \tag{47}$$

and

$$\phi \longleftarrow \phi - \alpha_\phi\nabla\phi. \tag{48}$$

Let α_w be the learning rate of w, then by the constraint $\mathrm{E}[X_T] = z$ we can get the standard stochastic approximation update rule:

$$w_{n+1} = w_n - \alpha_w\left(\frac{1}{m}\sum_{i=j-m+1}^{j} x_T^{(i)} - z\right),$$

where $x_T^{(i)}$ is the last point of sample i and $j \equiv 0 \mod m$.

We summarize the algorithm as pseudocode in Algorithm 1.

6. Numerical experiments

In this section, we first evaluate the performance of Algorithm 1 through simulations in section 6.1. We then test the proposed algorithm on real-world data in section 6.2, comparing it with traditional plug-in methods and entropy-regularized reinforcement learning algorithms in Wang and Zhou (2020).

6.1. *Simulation study*

In our setting, we consider an investment horizon of $T = 1$ year with a time step of $\Delta t = \frac{1}{252}$, corresponding to daily rebalancing over 252 trading days. The number of time grids is thus $N = 252$. We set the annualized risk-free rate to $r = 2\%$, while the annualized expected return μ and volatility σ are selected from $\{-50\%, -30\%, -10\%, 10\%, 30\%, 50\%\}$ and $\{10\%, 20\%, 30\%, 40\%\}$, respectively. Let the initial wealth to be $x_0 = 1$ and the annualized target return on the terminal wealth is 40% which yields $z = 1.4$.

For our algorithm, we take the number of episodes $K = 20000$, and take the sample average size for Lagrange multiplier $m = 10$. By Proposition 4.6 and Remark 4.8, to maintain exploration costs at comparable levels, we set the exploration weight λ to 0.01 when employing Φ_h as the regularizer, and 0.1 when using $\log\Phi_h$ as the regularizer. The learning rates are taken as $\alpha_\theta = \alpha_\phi = \alpha_w = 0.01$ with decay rate $l(j) = j^{-0.51}$.

Based on Examples 3.1 and 4.1, we mainly investigate the simulation results for three exploration distributions: Gaussian, exponential distribution and uniform distribution. We present the mean and the variance of the last 200 terminal wealth, and the corresponding Sharpe ratio ($\frac{mean-1}{\sqrt{variance}}$). The simulation results of our algorithm are presented in tables 1–3.

For different values of μ and σ, we take means of every 100 terminal wealth for different h to show the tendency of the expectation of terminal wealth in figures 1 and 2, respectively. Although different exploration preferences lead to different trajectories, the final outcomes are broadly similar: after sufficient iterations, the sample mean still fluctuates around 1.4. Moreover, the algorithm performs more significantly as $|\mu|$ increases or as σ decreases with other parameters fixed. In

Algorithm 1 Actor-Critic Algorithm for EMV Problem

Input: initial wealth x_0, the parameters (μ, σ, r) of Market, the target z, exploration weight λ, investment horizon T, time step Δt, number of time grids N, learning rates $\alpha_\theta,\ \alpha_\phi, \alpha_w$, number of episodes K, sample average size m, and a simulator of the market called *Market*.

Learning procedure: Initialize $\theta,\ \phi,\ w$.

for episode $j = 1$ **to** K **do**

Initialize $n = 0$

$x_{t_n} \leftarrow x_0$

while $n < N$ **do**

Compute and store $\frac{\partial}{\partial\theta} V^\theta(t_n, x_{t_n})$

Sample u_{t_n} from $\Pi^\phi(\cdot \mid t_n, x_{t_n})$.

Compute and store $p(t_n, x_{t_n}, \phi)$.

Compute and store $\frac{\partial p}{\partial \phi}(t_n, x_{t_n}, \phi)$.

Compute and store $\frac{\partial}{\partial\phi} \log \dot{\Pi}^\phi(u_{t_n} \mid t_n, x_{t_n})$.

Apply u_{t_n} to the market simulator and get the state x at next time point.

Store $x_{t_{n+1}} \leftarrow x$.

$n \leftarrow n + 1$.

end while

Store the terminal wealth $x_T^{(j)} \leftarrow x_{t_N}$.

Compute the gradient of θ and ϕ by (44) and (47), respectively.

$$\nabla\theta = -\sum_{i=0}^{N-1} \frac{\partial V^\theta}{\partial\theta}(t_i, x_{t_i})[V^\theta(t_{i+1}, x_{t_{i+1}}) - V^\theta(t_i, x_{t_i}) - \lambda p(t_i, x_{t_i}, \phi)\Delta t],$$

and

$$\nabla\phi = \sum_{i=0}^{N-1} \left\{ \frac{\partial \log \dot{\Pi}^\phi}{\partial\phi}(u_{t_i} \mid t_i, x_{t_i})[V^\theta(t_{i+1}, x_{t_{i+1}}) - V^\theta(t_i, x_{t_i}) - \lambda p(t_i, x_{t_i}, \phi)\Delta t] - \lambda \frac{\partial p}{\partial\phi}(t_i, x_{t_i}) \right\} + \textit{support adjustment term}.$$

Update θ and ϕ by (45) and (48), respectively.

$$\theta \longleftarrow \theta - \alpha_\theta l(j)\nabla\theta$$
$$\phi \longleftarrow \phi - \alpha_\phi l(j)\nabla\phi$$

Update w every m episodes:

if $j \equiv 0 \mod m$ **then**

$$w \leftarrow w - \alpha_w \left(\frac{1}{m} \sum_{i=j-m+1}^{j} x_T^{(i)} - z \right).$$

end if

end for

addition, when $|\mu|$ is small and σ is large relatively, the performance is bad. This is because larger σ reflects higher level of randomness of the environment, and at this time the significance of exploration becomes smaller.

The performance under different λ with Gaussian is shown in figures 3 and 4. We can see that when ρ^2 is relatively larger, λ has a more significant impact on algorithm performance under regularizer Φ_h than $\log \Phi_h$. This is consistent with Remark 4.8. Finally, we show one sample trajectory of u_{t_i} under different h in figure 5. As shown, the trajectories differ across regularizers: the exponential distribution exhibits a skewed exploration pattern, the normal distribution appears more symmetric, and the uniform distribution is more concentrated. Since our parameters and target settings are the same as those in Wang and Zhou (2020), we can see that our RL algorithm based on Choquet regularizations and logarithmic Choquet regularizers perform on par with the one in Wang and Zhou (2020). Compared with the results that Gaussian is always the optimal in Wang and Zhou (2020), the availability of a large class of Choquet regularizers makes it possible

Table 1. Performance of Gaussian with $h(p) = \int_0^p z(1-s)\,\mathrm{d}s$.

		Φ_h			$\log \Phi_h$		
μ	σ	Mean	Variance	Sharpe ratio	Mean	Variance	Sharpe ratio
−0.5	0.1	1.4052	0.0035	6.8192	1.4052	0.0037	6.6520
−0.3	0.1	1.4141	0.0103	4.0852	1.4143	0.0104	4.0554
−0.1	0.1	1.4479	0.1104	1.3482	1.4485	0.1107	1.3482
0.1	0.1	1.3966	0.2516	0.7906	1.3970	0.2571	0.7828
0.3	0.1	1.4052	0.0408	2.0043	1.4055	0.0441	1.9307
0.5	0.1	1.4007	0.0247	2.5722	1.4007	0.0267	2.4519
−0.5	0.2	1.4078	0.0147	3.3654	1.4077	0.0153	3.2939
−0.3	0.2	1.4208	0.0458	1.9668	1.4209	0.0464	1.9534
−0.1	0.2	1.4557	0.5046	0.6416	1.4552	0.5038	0.6413
0.1	0.2	1.3576	0.8506	0.3878	1.3575	0.8643	0.3846
0.3	0.2	1.3967	0.1402	1.0595	1.3966	0.1487	1.0284
0.5	0.2	1.3943	0.0739	1.4506	1.3941	0.0799	1.3945
−0.5	0.3	1.4118	0.0368	2.1456	1.4117	0.0382	2.1053
−0.3	0.3	1.4290	0.1201	1.2362	1.4292	0.1221	1.2282
−0.1	0.3	1.4143	1.0305	0.4081	1.4126	1.0228	0.4080
0.1	0.3	1.2978	1.3627	0.2551	1.2974	1.3796	0.2532
0.3	0.3	1.3887	0.2825	0.7314	1.3884	0.2961	0.7138
0.5	0.3	1.3890	0.1353	1.0574	1.3886	0.1444	1.0225
−0.5	0.4	1.4171	0.0761	1.5122	1.4169	0.0786	1.4872
−0.3	0.4	1.4364	0.2507	0.8715	1.4366	0.2539	0.8665
−0.1	0.4	1.3539	1.4238	0.2966	1.3514	1.4054	0.2965
0.1	0.4	1.2358	1.5370	0.1902	1.2346	1.5465	0.1887
0.3	0.4	1.3801	0.4691	0.5550	1.3797	0.4879	0.5436
0.5	0.4	1.3844	0.2119	0.8351	1.3839	0.2244	0.8103

Table 2. Performance of exponential distribution with $h(p) = -p\log p$.

		Φ_h			$\log \Phi_h$		
μ	σ	Mean	Variance	Sharpe ratio	Mean	Variance	Sharpe ratio
−0.5	0.1	1.4081	0.0041	6.3781	1.4105	0.0049	5.8661
−0.3	0.1	1.4194	0.0109	4.0208	1.4200	0.0132	3.6508
−0.1	0.1	1.4617	0.1069	1.4119	1.4531	0.1249	1.2822
0.1	0.1	1.3743	0.2063	0.8240	1.3761	0.2493	0.7533
0.3	0.1	1.3979	0.0372	2.0637	1.4039	0.0387	2.0543
0.5	0.1	1.3970	0.0229	2.6243	1.4018	0.0232	2.6370
−0.5	0.2	1.4144	0.0165	3.2244	1.4150	0.0199	2.9382
−0.3	0.2	1.4332	0.0475	1.9866	1.4271	0.0571	1.7874
−0.1	0.2	1.4821	0.4784	0.6970	1.4551	0.5237	0.6289
0.1	0.2	1.3262	0.6552	0.4031	1.3167	0.7928	0.3557
0.3	0.2	1.3853	0.1246	1.0915	1.3934	0.1278	1.1006
0.5	0.2	1.3880	0.0696	1.4713	1.3983	0.0697	1.5090
−0.5	0.3	1.4228	0.0396	2.1254	1.4184	0.0474	1.9159
−0.3	0.3	1.4503	0.1213	1.2929	1.4318	0.1409	1.1504
−0.1	0.3	1.4243	0.8750	0.4536	1.4212	1.0341	0.4143
0.1	0.3	1.2595	0.9949	0.2602	1.2460	1.2313	0.2217
0.3	0.3	1.3767	0.2393	0.7700	1.3802	0.2413	0.7739
0.5	0.3	1.3818	0.1239	1.0847	1.3923	0.1223	1.1219
−0.5	0.4	1.4334	0.0771	1.5610	1.4207	0.0914	1.3916
−0.3	0.4	1.4671	0.2474	0.9390	1.4344	0.2741	0.8298
−0.1	0.4	1.2964	0.8256	0.3263	1.3428	1.1752	0.3162
0.1	0.4	1.1885	1.0576	0.1833	1.1782	1.3137	0.1555
0.3	0.4	1.3691	0.3772	0.6010	1.3656	0.3720	0.5994
0.5	0.4	1.3777	0.1872	0.8729	1.3849	0.1797	0.9080

to choose specific regularizers to achieve certain objective used exploratory samplers such as exponential, uniform and Gaussian.

6.2. Real data analysis

Unlike simulations, empirical analysis requires more practical considerations and methodological adjustments. Following the approach in Huang *et al.* (2022), we set up our empirical analysis as follows. First, since the trading occurs after we observe the stock price at the trading moment, we cannot update the parameters after observing the data for a whole year. Thus, the algorithm must be adapted for online updating to reflect real-time decision-making. Second, exploration increases variance, and since wealth dynamics are fully determined by observed stock prices, we can track wealth

Table 3. Performance of uniform distribution with $h(p) = p - p^2$.

		Φ_h			$\log \Phi_h$		
μ	σ	Mean	Variance	Sharpe ratio	Mean	Variance	Sharpe ratio
−0.5	0.1	1.4100	0.0028	7.8029	1.4100	0.0033	7.1396
−0.3	0.1	1.4214	0.0093	4.3755	1.4197	0.0111	3.9871
−0.1	0.1	1.4632	0.1042	1.4352	1.4530	0.1203	1.3059
0.1	0.1	1.3859	0.1848	0.8977	1.3744	0.2374	0.7684
0.3	0.1	1.4049	0.0204	2.8296	1.4033	0.0243	2.5886
0.5	0.1	1.4015	0.0101	3.9945	1.4010	0.0113	3.7791
−0.5	0.2	1.4171	0.0117	3.8522	1.4145	0.0140	3.4981
−0.3	0.2	1.4361	0.0414	2.1422	1.4267	0.0485	1.9384
−0.1	0.2	1.4834	0.4646	0.7093	1.4548	0.5042	0.6405
0.1	0.2	1.3493	0.6026	0.4499	1.3115	0.7657	0.3560
0.3	0.2	1.4007	0.0799	1.4174	1.3911	0.0935	1.2786
0.5	0.2	1.3995	0.0370	2.0771	1.3958	0.0415	1.9428
−0.5	0.3	1.4260	0.0292	2.4917	1.4180	0.0345	2.2490
−0.3	0.3	1.4537	0.1079	1.3812	1.4317	0.1209	1.2416
−0.1	0.3	1.4225	0.8272	0.4645	1.4203	0.9898	0.4225
0.1	0.3	1.2918	0.9498	0.2994	1.2374	1.1943	0.2172
0.3	0.3	1.3985	0.1768	0.9477	1.3762	0.2017	0.8377
0.5	0.3	1.3987	0.0785	1.4232	1.3883	0.0868	1.3180
−0.5	0.4	1.4367	0.0593	1.7931	1.4207	0.0680	1.6135
−0.3	0.4	1.4708	0.2236	0.9955	1.4347	0.2376	0.8918
−0.1	0.4	1.2928	0.7377	0.3409	1.3401	1.1045	0.3235
0.1	0.4	1.2258	1.0429	0.2211	1.1673	1.2627	0.1488
0.3	0.4	1.3957	0.3079	0.7131	1.3598	0.3418	0.6153
0.5	0.4	1.3986	0.1339	1.0891	1.3794	0.1449	0.9967

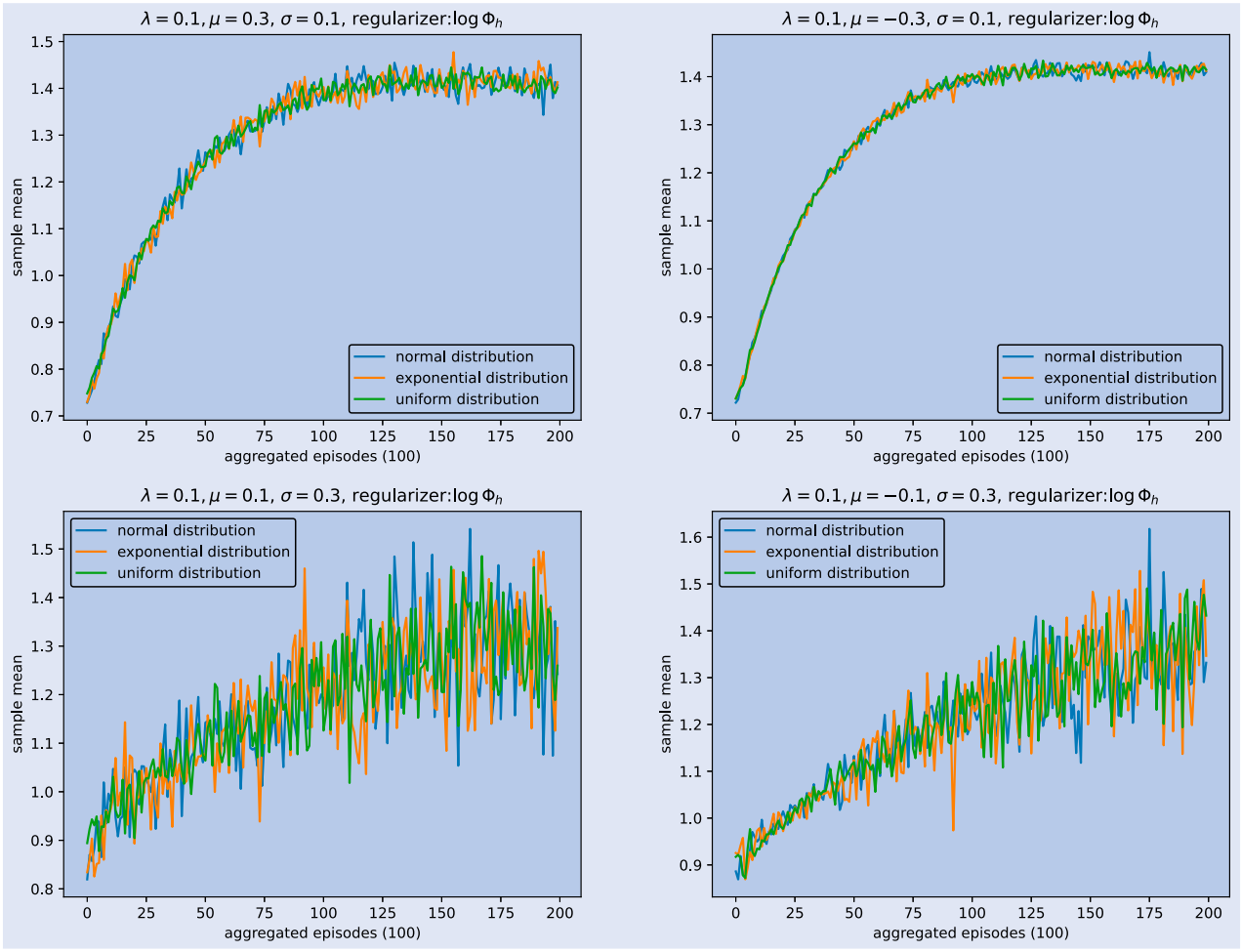

Figure 1. The effect of μ and σ on the exploration for the regularizer $\log \Phi_h$.

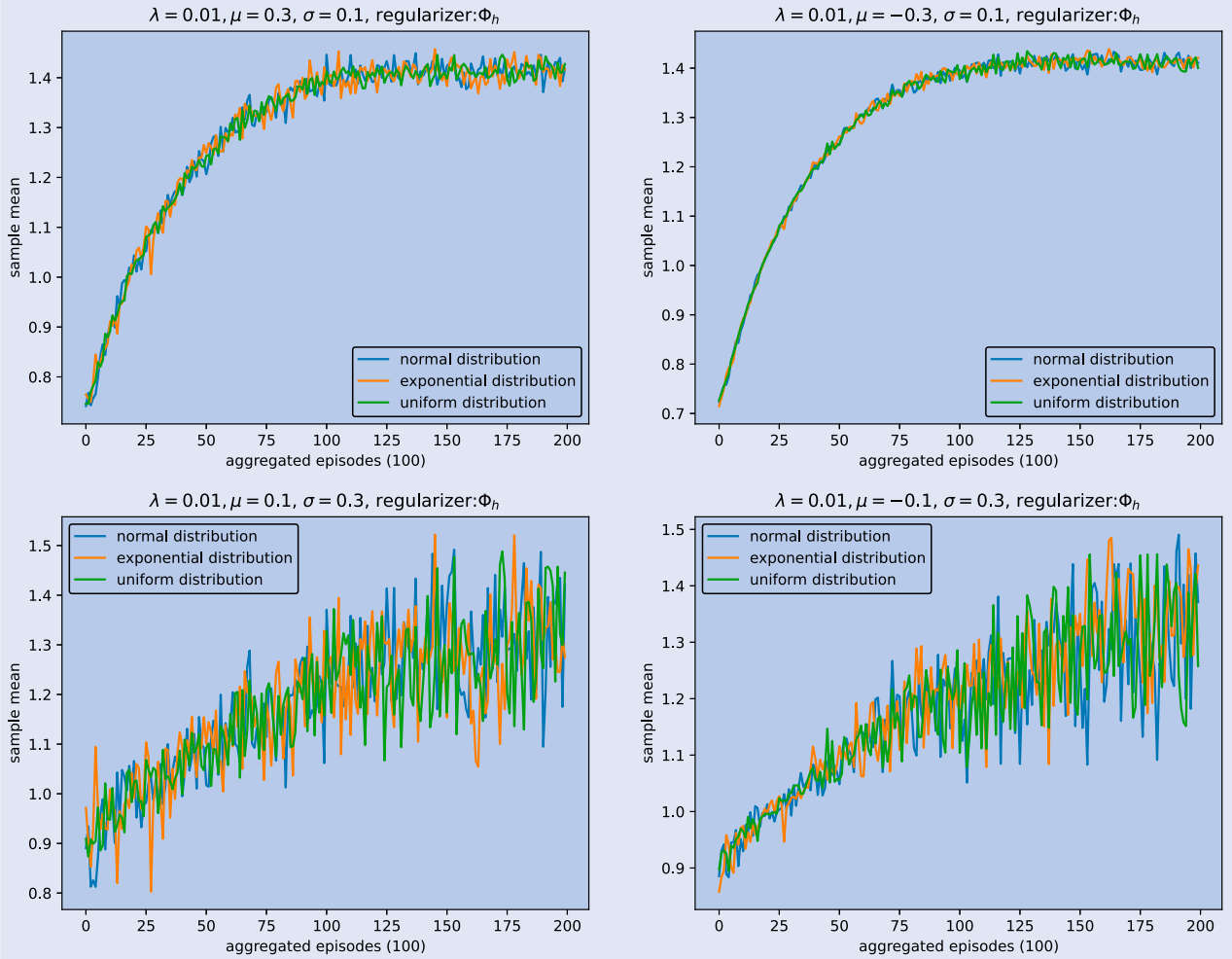

Figure 2. The effect of μ and σ on the exploration for the regularizer Φ_h.

evolution without actually executing any specific investment policy. Gradient estimation requires stochastic policies, but the expected value of these policies corresponds to the optimal deterministic policy at each point in time. Therefore, we train using stochastic policies and execute investments using their deterministic counterparts. This constitutes a form of off-policy learning. Third, real financial markets are subject to various sources of data instability, such as regime shifts, natural disasters, and economic fluctuations. Consequently, we avoid techniques like sample reuse, which are effective in controlled simulations but may introduce bias in empirical settings. Instead, we employ mini-batch updates to reduce the variance in gradient estimates. Finally, prior to the backtesting phase, we perform pre-training using historical data from before the test period to provide a stable initial model.

We obtain daily adjusted closing prices for 50 U.S. equities spanning the period from 1990 to 2019 via the Yahoo Finance API using Python. These publicly available data are used exclusively for academic research. Among them, the data from 1990 to 1999 is used for pre-training, and the data from 2000 to 2019 is used for backtesting. Each time we select one stock and repeat the experiment 50 times to collect statistics. In each experiment, the initial wealth is set to 1, and the investment horizen is fixed at one year ($T = 1$). We then conducted continuous backtesting for 20 years. The S&P 500 Index began at a level of 359.69 at the start of 1990 and rose to 1464.47 by the end of 1999. The annualized return over this 10-year period is calculated using the compound annual growth rate (CAGR) formula:

$$\text{CAGR} = \left(\frac{1464.47}{359.69}\right)^{\frac{1}{10}} - 1 \approx 0.1507.$$

This implies that the S&P 500 Index achieved an approximate annualized return of 15.07% from 1990 to 1999, and we accordingly set the target return to 15%. In our model, investors are assumed to be price takers. Taking into account practical considerations such as taxes and transaction costs, low-frequency trading is more appropriate. Therefore, we adopt a monthly rebalancing schedule.

We compare the performance of three different forms of the distortion function h under the Choquet regularizer, as well as the performance of the logarithmic Choquet regularizer. In addition, we benchmark these results against the classical plug-in policy. As noted in Remark 4.3, the logarithmic Choquet regularizer encompasses Shannon entropy as a special case under a specific choice of the distortion function. Consequently, under normal exploration, the results obtained using the log-Choquet regularizer correspond to the results using entropy regularizer in Wang and Zhou (2020).

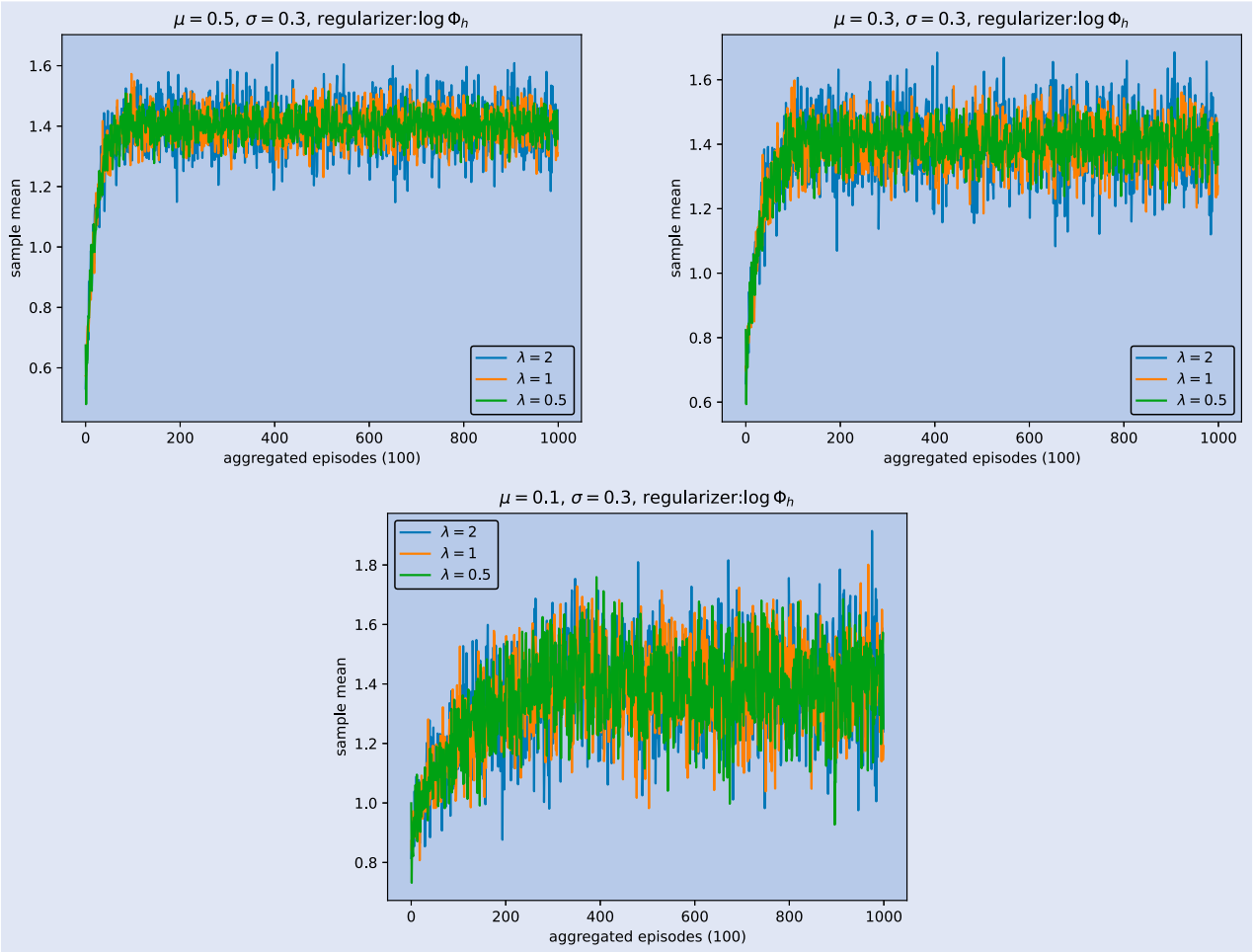

Figure 3. The effect of λ on the exploration for the regularizer $\log \Phi_h$.

Following Huang *et al.* (2022), we first consider the case with a leverage constraint, where borrowing is not permitted. We also explore the scenario without this constraint. The average wealth trajectories with leverage constraint under different regularizers are reported in figures 6 and 7. And the average wealth trajectories without leverage constraint under different regularizers are reported in figures 8 and 9. Note that the plug-in method is excessively sensitive to historical data, particularly in the presence of extreme values. This often results in overly aggressive strategies that, without leverage constraints, produce highly volatile wealth trajectories. As these trajectories offer limited interpretive value, we omit them from the figures. Summary statistics, including the annualized return (AR), the annualized volatility (AV) and the Sharpe ratio (SR) are reported in tables 4 and 5.

Based on the figures and tables, we summarize our main findings as follows. First, by comparing figures 6 and 7, or figures 8 and 9, we observe that choosing the log-form Choquet regularization or using the non-log version has minimal impact on the algorithm's performance. The results under both types of regularizers are highly similar, which is consistent with our simulation findings. This observation is further supported by the summary statistics in tables 4 and 5, where the performance metrics under the two regularizers are nearly identical when the same exploration distribution is used.

Second, unlike the simulation study in section 6.1 where the stock price has a fixed growth trend, when leverage constraints are absent, exploration based on the exponential distribution leads to slightly faster wealth accumulation after the financial crisis and a higher average terminal wealth. This behavior is attributable to the asymmetric and heavy-tailed characteristics of the exponential distribution, which allows for broader exploration at larger investment levels. When the stock price shifts from a decline to an increase, this leads to faster wealth growth. However, as shown in figures 8 and 9, it also lead to a sharper decline in wealth under adverse market conditions, such as the financial crisis. Since the result obtained with the logarithmic Choquet regularizer correspond to the result with entropy regularizer in Wang and Zhou (2020), by adopting alternative forms of the distortion function h, our Choquet regularization framework exhibits more flexibility. From this perspective, the Choquet regularizer demonstrates superior robustness and adaptability compared to Shannon's differential entropy.

Third, the limited exploration range of the uniform distribution, as well as the lower bound inherent in the exponential distribution, result in faster wealth accumulation compared to the normal distribution. This is because negative investment exploration reduces returns during economic upswings. This behavior is also evident in the empirical analysis: during the pronounced bull market from 2010 to 2020, exploration

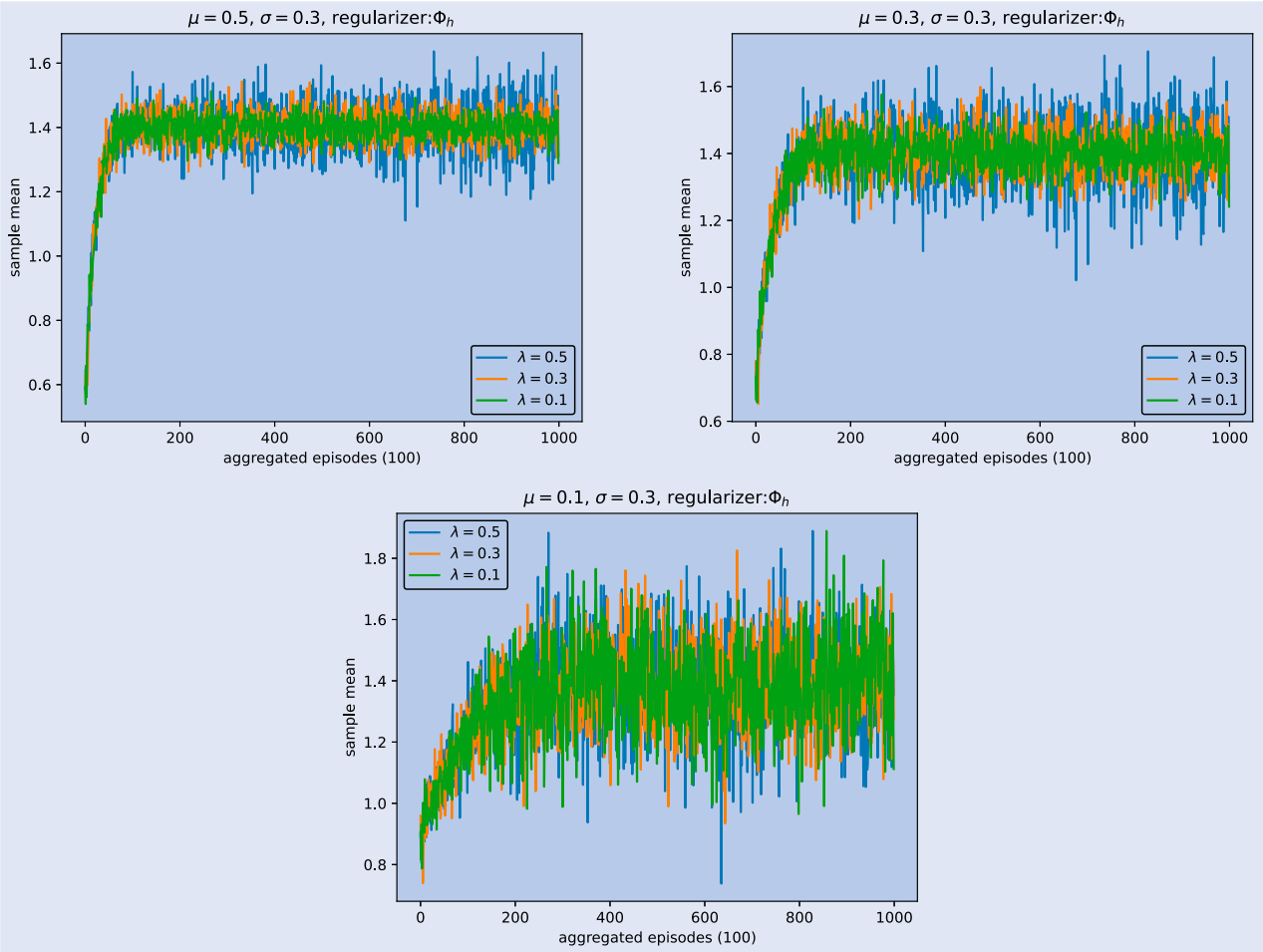

Figure 4. The effect of λ on the exploration for the regularizer Φ_h.

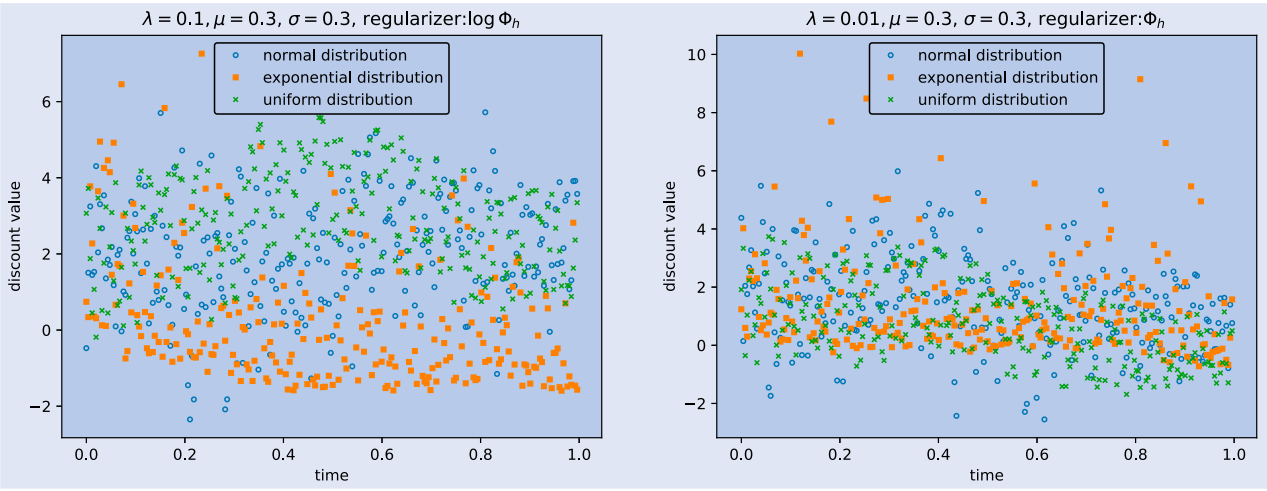

Figure 5. Samples of u_{t_i} for the regularizer log Φ_h and Φ_h.

based on the exponential distribution and uniform distribution exhibited faster growth in wealth.

Finally, after imposing the leverage constraint, the exploration policy based on the exponential distribution exhibits the most pronounced change, reflecting the heavier tail characteristics of the exponential distribution compared to others. Despite this adjustment, its overall performance remains superior to that of the plug-in method, thereby demonstrating the robustness of our algorithm.

7. Conclusion

For the first time, we applied the Choquet-regularized continuous-time RL framework proposed by Han *et al.* (2023) to practical problems. We studied the MV problem under Choquet regularization and its logarithmic form. Several different optimal exploration distributions of different h were given, and when $\|h'\|_2$ is fixed, the optimal exploration distributions have the same mean and variance. Unlike the infinite time

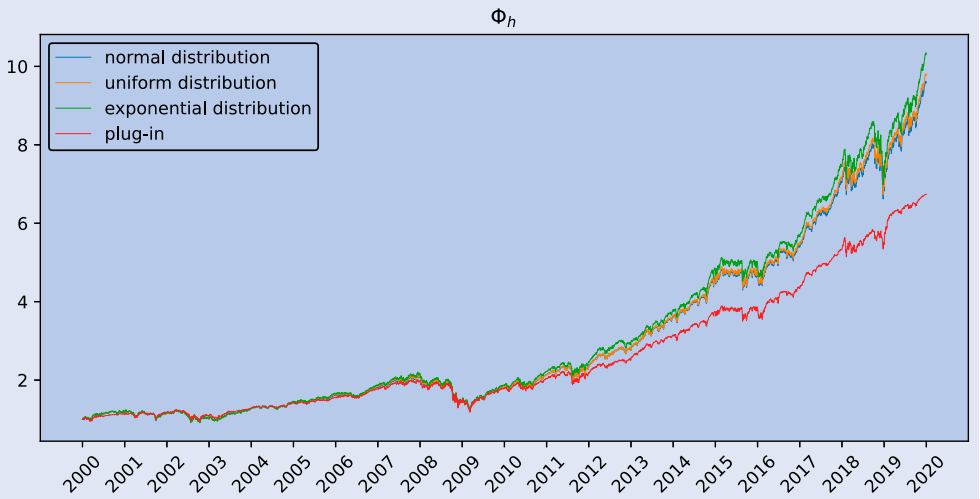

Figure 6. The average wealth trajectory under Φ_h with leverage constraint from 2000 to 2020.

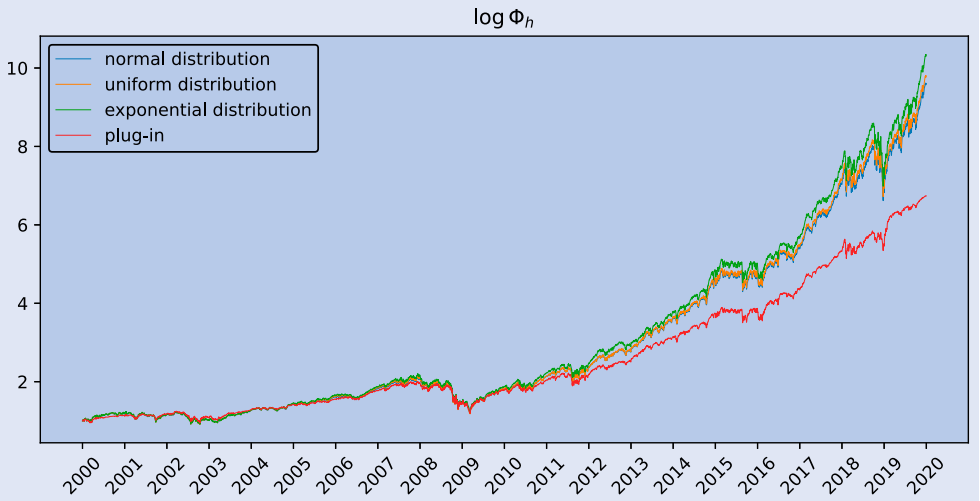

Figure 7. The average wealth trajectory under $\log \Phi_h$ with leverage constraint from 2000 to 2020.

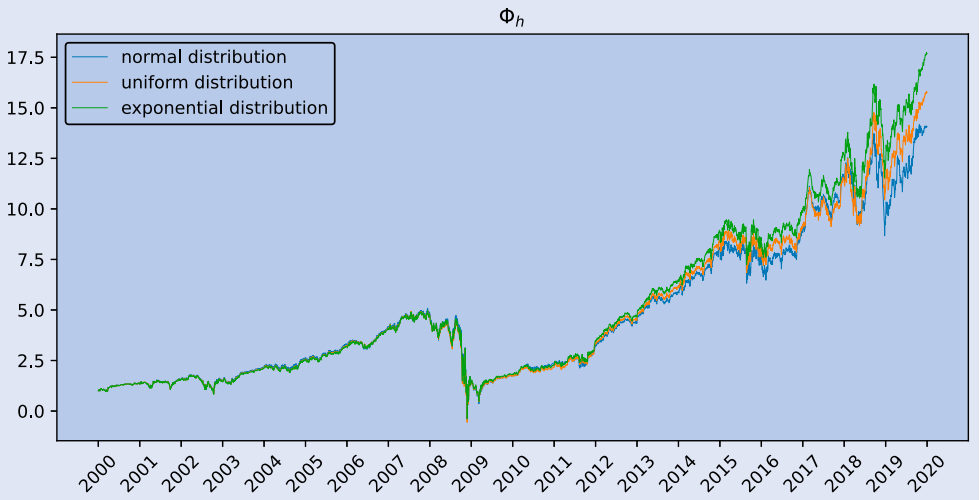

Figure 8. The average wealth trajectory under Φ_h without leverage constraint from 2000 to 2020.

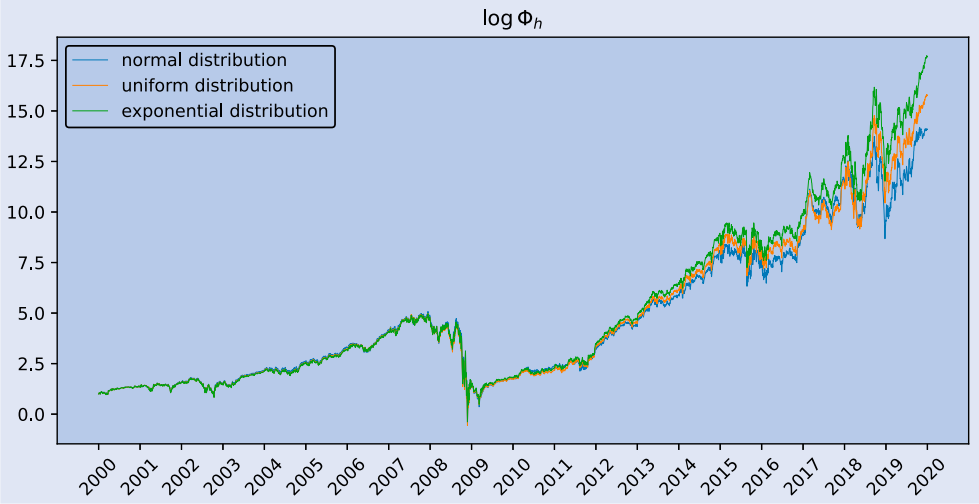

Figure 9. The average wealth trajectory under log Φ_h without leverage constraint from 2000 to 2020.

Table 4. Performance with leverage constraint.

		Φ_h			log Φ_h		
	Plug-in	Normal	Uniform	Exponential	Normal	Uniform	Exponential
AR	0.0913	0.1136	0.1149	0.1182	0.1132	0.1149	0.1151
AV	0.1407	0.1590	0.1630	0.1628	0.1597	0.1630	0.1634
SR	0.6487	0.7146	0.7049	0.7260	0.7090	0.7049	0.7043

Table 5. Performance without leverage constraint.

	Φ_h			log Φ_h		
	Normal	Uniform	Exponential	Normal	Uniform	Exponential
AR	0.1372	0.1441	0.1511	0.1373	0.1441	0.1511
AV	1.4240	2.4765	1.2660	1.4186	2.4801	0.2655
SR	0.0963	0.0582	0.1193	0.0968	0.0581	0.1194

horizon results in Han *et al.* (2023), the variance decreases over time in the finite time horizon problem. At the same time, the mean of the optimal exploration distribution is related to the current state x and independent of λ and h, which is equal to the optimal action of the classical MV problem. The variance of the optimal exploration distribution is related to λ and h and independent of state x, and even independent of h under logarithmic regularization. These also showed the perfect separation between exploitation and exploration in the mean and variance of the optimal distributions as in Wang and Zhou (2020) when entropy is used as a regularizer.

Further, we have obtained that the two regularization problems converge to the traditional MV problem, and compared the exploration costs of the two regularizations. We found that the exploration cost under the logarithmic Choquet regularization is consistent with the exploration cost under the entropy regularization, only related to λ and time range T, while the exploration cost under Choquet regularization is also related to market parameters. Through simulation, we compared the two kinds of regularization. In general, when the market fluctuates greatly and the willingness to explore is not strong, the cost of Choquet regularization is lower. On the contrary, it may be better to use logarithmic Choquet regularizers for regularization.

There remain several open questions to address. Firstly, we currently treat λ as an exogenous variable. However, from the perspective of exploration cost, turning λ into endogenous and changeable can help us better control the exploration cost. As time goes by, the information we obtain through exploration will also increase, so the willingness to explore will also change, which also implies the rationality of the changing λ to time-related. Secondly, investigating stochastic differential games within the mean-variance framework using reinforcement learning presents an intriguing avenue for exploration. Thirdly, the current Choquet integral is limited to handling one-dimensional action spaces. Extending Choquet regularization to high-dimensional settings remains an open challenge. As discussed in Han *et al.* (2023), joint and marginal-based constructions offer two promising directions, and a comprehensive theoretical study along these lines is left for future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research of Junyi Guo is supported by the National Natural Science Foundation of China (Grant No. 12271274). The research of Xia Han is supported by the National Natural Science Foundation of China (Grant Nos. 12301604,12371471, 12471449).

References

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

Doya, K., Reinforcement learning in continuous time and space. *Neural. Comput.*, 2000, **12**(1), 219–245.

Gilboa, I. and Schmeidler, D., Maxmin expected utility with non-unique prior. *J. Math. Econ.*, 1989, **18**(2), 141–153.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R.E. and Levine, S., Q-prop: Sample-efficient policy gradient with an off-policy critic, 2016. arXiv: 1611.02247.

Guo, X., Xu, R. and Zariphopoulou, T., Entropy regularization for mean field games with learning. *Math. Oper. Res.*, 2022, **47**(4), 3239–3260.

Haarnoja, T., Tang, H., Abbeel, P. and Levine, S., Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1352–1361, 2017 (PMLR: Sydney).

Han, X., Wang, R. and Zhou, X.Y., Choquet regularization for continuous-time reinforcement learning. *SIAM J. Control Optim.*, 2023, **61**(5), 2777–2801.

Hu, T. and Chen, O., On a family of coherent measures of variability. *Insurance: Math. Econ.*, 2020, **95**, 173–182.

Huang, Y., Jia, Y. and Zhou, X., Achieving mean-variance efficiency by continuous-time reinforcement learning. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 377–385, 2022 (Association for Computing Machinery: New York, NY).

Jia, Y. and Zhou, X.Y., Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *J. Mach. Learn. Res.*, 2022a, **23**(154), 1–55.

Jia, Y. and Zhou, X.Y., Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *J. Mach. Learn. Res.*, 2022b, **23**(275), 1–50.

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

Konda, V. and Tsitsiklis, J, Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pp. 1008–1014, 1999 (MIT Press: Cambridge, MA).

Li, D. and Ng, W.L., Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Math. Finance*, 2000, **10**(3), 387–406.

Li, X., Zhou, X.Y. and Lim, A.E., Dynamic mean-variance portfolio selection with no-shorting constraints. *SIAM J. Control Optim.*, 2002, **40**(5), 1540–1555.

Liu, F., Cai, J., Lemieux, C. and Wang, R., Convex risk functionals: Representation and applications. *Insurance: Math. Econ.*, 2020, **90**, 66–79.

Markowitz, H., Portfolio selection. *J. Finance.*, 1952, **7**(1), 77–91.

Neu, G., Jonsson, A. and Gómez, V., A unified view of entropy-regularized markov decision processes, 2017. arXiv: 1705.07798.

Pesenti, S., Wang, Q. and Wang, R., Optimizing distortion riskmetrics with distributional uncertainty. *Math. Program.*, 2025, **213**, 51–106.

Quiggin, J., A theory of anticipated utility. *J. Econ. Behavior Organ.*, 1982, **3**(4), 323–343.

Rao, M., Chen, Y., Vemuri, B.C. and Wang, F., Cumulative residual entropy: A new measure of information. *IEEE Trans. Inf. Theory*, 2004, **50**(6), 1220–1228.

Sunoj, S.M. and Sankaran, P.G., Quantile based entropy function. *Stat. Probab. Lett.*, 2012, **82**(6), 1049–1053.

Sutton, R.S. and Barto, A.G., *Reinforcement Learning: An Introduction*, 2018 (MIT Press: Cambridge, MA).

Wang, H., Zariphopoulou, T. and Zhou, X.Y., Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 2020a, **21**(198), 1–34.

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

Wang, R., Wei, Y. and Willmot, G.E., Characterization, robustness and aggregation of signed Choquet integrals. *Math. Oper. Res.*, 2020b, **45**(3), 993–1015.

Zhou, X.Y. and Li, D., Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Appl. Math. Optim.*, 2000, **42**(4), 19–33.

Zhou, X.Y, Curse of optimality, and how do we break it, 2021. SSRN: 3845462.

Ziebart, B.D., Modeling purposeful adaptive behavior with the principle of maximum causal entropy. PhD Thesis, 2010.