Continuous-Time Mean–Variance Portfolio Optimization via Reinforcement Learning

Haoran Wang*  Xun Yu Zhou†

First draft: February 2019

Abstract

We consider continuous-time Mean–variance (MV) portfolio optimization problem in the Reinforcement Learning (RL) setting. The problem falls into the entropy-regularized relaxed stochastic control framework recently introduced in Wang et al. (2019). We derive the feedback exploration policy as the Gaussian distribution, with time-decaying variance. Close connections between the entropy-regularized MV and the classical MV are also discussed, including the solvability equivalence and the convergence as exploration decays. Finally, we prove a policy improvement theorem (PIT) for the continuous-time MV problem under both entropy regularization and control relaxation. The PIT leads to an implementable RL algorithm for the continuous-time MV problem. Our algorithm outperforms an adaptive control based method that estimates the underlying parameters in real-time and a state-of-the-art RL method that uses deep neural networks for continuous control problems by a large margin in nearly all simulations.

Key words. Reinforcement learning, entropy regularization, Mean-variance portfolio optimization, stochastic control, Gaussian distribution.

*Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. Email: hw2718@columbia.edu.

†Department of Industrial Engineering and Operations Research, and The Data Science Institute, Columbia University, New York, NY 10027, USA. Email: xz2574@columbia.edu.
1 Introduction

The application of Reinforcement Learning (RL) to quantitative finance (e.g., algorithmic and high frequency trading, smart order routing, portfolio management, etc) has gained more attention in recent years. One of the main reasons is that the electronic markets prevailing nowadays can provide sufficient amount of microstructure data for training and adaptive learning beyond the handling capacity of human traders and portfolio managers. Various efforts have been made in this direction; for example, Nevmyvaka et al. (2006) conducted the first large scale empirical analysis of RL method applied to optimal order execution and achieved substantial improvement relative to baseline strategies. Hendricks and Wilcox (2014) improved over the theoretical optimal trading trajectory in the Almgren-Chriss model (Almgren and Chriss (2001)) using RL techniques and market attributes. Moody and Saffell (2001), Moody et al. (1998) studied portfolio allocation problems by taking into account transaction costs via direct policy search based RL methods, without resorting to forecast methods that rely on supervised leaning. However, most existing works only focus on risk-neutral RL optimization problems under the expectation criterion. Such criterion cannot fully characterize the uncertainty of the decision making process in the financial markets, where uncertainty, commonly represented by the variance of the payoff, is crucial for risk-averse investment and trading in practice.

Mean–variance (MV) criterion is one of the most important criteria for optimization problems in quantitative finance. Initiated in the seminal work Markowitz (1952) for portfolio selection problem in a single period, such criterion yields an asset allocation strategy that minimizes the variance of the payoff while targeting some prespecified mean return. The MV problem has been further investigated in discrete time multiperiod setting (Li and Ng (2000)) and continuous time (Zhou and Li (2000)), and also studied for hedging problems (Duffie and Richardson (1991)) and optimal liquidation problems (Almgren and Chriss (2001)), among various generalizations. The popularity of the MV criterion is not only due to its intuition and simplicity in capturing both risk and reward for practitioners, but also due to the academically intriguing issue of time-inconsistency (or Bellman’s inconsistency) in the underlying stochastic optimization and control problems. Also, from the RL perspective, as pointed out in Mannor and Tsitsiklis (2013), it is computationally challenging to do planning and seek the global optimum for Markov Decision Process (MDP) problems under the MV criterion.

In fact, variance estimation and control is not as direct as optimizing the
expected reward-to-go which is well understood in the classical MDP framework for most RL problems. Due to the fact that most of the standard performance criteria for MDP are linear in expectation, including the discounted sum of rewards and the average reward settings (Sutton and Barto (2018)), Bellman’s consistency equation can be easily derived for guiding policy evaluation or control and has led to many state-of-the-art RL methods (e.g., Q-learning, temporal difference (TD) learning, etc). The variance of reward-to-go, however, is nonlinear in expectation and, as a result, most of the well-known learning rules cannot be applied directly. Existing works on variance estimation and control generally divide into two groups, value based methods and policy based methods. Sobel (1982) obtained the Bellman equation for the variance of reward-to-go given a fixed policy, which is nonlinear in the expected reward-to-go, and from there, Sato et al. (2001) derived the TD(0) learning rule to estimate the variance under a given policy. In the related paper Sato and Kobayashi (2000), this value based method was applied to an MV control problem in the portfolio selection setting. It is worth noting, however, due to their definition of the intermediate value function (i.e., variance penalized expected reward-to-go), the Bellman’s optimality equation does not hold. It hence cannot guarantee that a greedy policy with respect to the latest updated value function would eventually lead to the optimal policy. A second direction that uses policy based RL methods was suggested in Tamar et al. (2013). They also extended the work to linear function approximators and actor-critic algorithm for the MV optimization problem for which convergence to the local optimum is guaranteed with probability 1 (Tamar and Mannor (2013)). Related works following this line of research include Prashanth and Ghavamzadeh (2013, 2016), among others. Despite the various methods mentioned above, it remains to be an open and interesting question to search for the global optimum under the MV criterion using RL methods.

In this paper, we propose to study the continuous time MV portfolio optimization problem, under the entropy-regularized relaxed stochastic control framework in Wang et al. (2019) recently developed by the authors and their collaborator. The continuous time formulation is specifically appealing when the rebalancing of the portfolio happens at ultra-high frequency. It can also benefit from the large amount of microstructure data that is relatively easy to obtain in most electronic markets today. Moreover, the readily available tools including stochastic calculus and differential equations in continuous time enable us to derive elegant and insightful results, which turns out to be essential in algorithm design for discrete time implementation (see section 4). For studying RL problems in continuous time, the
work [Wang et al. (2019)] proposed to explicitly capture exploration in both the reward function and the state dynamics, leading to non-Dirac valued control distributions in general. Specifically, the stochastic linear–quadratic (LQ) problem has been extensively studied in [Wang et al. (2019)]. Being essentially a special case of the LQ problems, the MV problem can be reformulated in its “exploratory” version under the entropy-regularized relaxed stochastic control framework and, it is the first contribution of this paper to present the global optimal solution to the exploratory MV problem. One of the interesting findings is that, unlike the infinite horizon exploratory LQ control case in [Wang et al. (2019)], the optimal feedback control distribution for the finite horizon exploratory MV problem is a Gaussian distribution with time-decaying variance. Similar to the LQ case in [Wang et al. (2019)], we also observed herein the perfect separation between exploitation and exploration in the mean and variance of the Gaussian distribution, as well as the effect on the exploration presented by the learning opportunity from the underlying environment. Another contribution is that we established close connections between the classical MV problem and the exploratory MV problem, including the solvability equivalence and the convergence result as exploration decays. The exploration cost for the finite horizon exploratory MV problem also turns out to be a constant that resembles the cost for the infinite horizon exploratory LQ case in [Wang et al. (2019)].

We then design an implementable RL algorithm to learn the global optimal solution of the exploratory MV problem, based on a provable policy improvement theorem for continuous time stochastic control problems with both entropy regularization and control relaxation. This theorem provides an explicit updating scheme of the feedback Gaussian policy, based on the value function of the current policy in an iterative manner. Its practical importance also includes the reduction from a family of general non-parametric policies to a specifically parametrized Gaussian family for exploration and exploitation, irrespective of the starting exploration policy. This, together with a carefully chosen initial Gaussian policy at the beginning of the learning process (see Theorem 6), guarantee the fast convergence of both the policy and the value function to the global optimum of the exploratory MV problem. We further compare our RL algorithm with two other methods in the MV portfolio optimization setting. The first approach is an adaptive control approach that adopts the real-time estimation of the underlying model parameters using maximum likelihood method. We also compare our RL algorithm to a state-of-the-art continuous control RL algorithm, the deep deterministic policy gradient method [Lillicrap et al. (2016)] that uses deep neural networks. The comparisons are performed under various sim-
ulated market scenarios, including both the stationary and non-stationary investment opportunities cases. Across nearly all the simulations, our RL algorithm outperforms the other two methods by a large margin, in terms of both performance and computing time.

The rest of the paper is organized as follows. In section 2, we present the continuous time exploratory MV problem under the entropy-regularized relaxed stochastic control framework introduced in Wang et al. (2019). Section 3 provides the complete solution of the exploratory MV problem, with connections to its classical counterpart also being established. We then provide the policy improvement theorem and a convergence result for the learning problem in section 4, based on which we introduce the RL algorithm for solving the exploratory MV problem. In section 5, we compare our method with two other alternative methods by performing simulations under various market scenarios. Finally, we conclude in section 6 and discuss some possible future research directions.

2 Formulation of Problem

In this section, we formulate an exploratory, entropy-regularized Markowitz’s MV portfolio selection problem in continuous time, in the context of RL. The motivation of a general exploratory stochastic control formulation, of which the MV problem is a special case, was discussed at great length in a previous paper Wang et al. (2019); so we will frequently refer to that paper.

2.1 Classical continuous-time MV problem

We first recall the classical MV problem in continuous time (without RL). For ease of presentation, throughout this paper we consider an investment universe consisting of only one risky asset and one riskless asset. The case of multiple risky assets poses no essential differences or difficulties other than notational complexity.

Let an investment horizon $T > 0$ be fixed, and $\{W_t, 0 \leq t \leq T\}$ a standard one-dimensional Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0 \leq t \leq T}, \mathbb{P})$ that satisfies the usual conditions. The price process of the risky asset is a geometric Brownian motion:

$$dS_t = S_t (\mu dt + \sigma dW_t), \quad 0 \leq t \leq T,$$

with $S_0 = s > 0$ being the initial price at $t = 0$, and $\mu \in \mathbb{R}, \sigma > 0$ being the mean and volatility parameters, respectively. The riskless asset has a
constant interest rate $r > 0$. The Sharpe ratio of the risky asset is defined by $\rho = \frac{\mu - r}{\sigma}$.

Denote by $\{x_u^t, 0 \leq t \leq T\}$ the discounted wealth process of an agent who rebalances her portfolio investing in the risky and riskless assets with a strategy $u = \{u_t, 0 \leq t \leq T\}$. Here $u_t$ is the discounted dollar value put in the risky asset at time $t$, while satisfying the standard self-financing assumption and other technical conditions that will be spelled out in details below. It follows from (1) that the wealth process satisfies

$$dx_u^t = \sigma u_t (\rho \, dt + dW_t), \quad 0 \leq t \leq T,$$

with an initial endowment being $x_u^0 = x \in \mathbb{R}$.

The classical continuous-time MV model aims to solve the following constrained optimization problem

$$\min_{u} \text{Var}[x_u^T]$$

subject to $\mathbb{E}[x_u^T] = z,$

where $\{x_u^t, 0 \leq t \leq T\}$ satisfies the dynamics (2) and $z \in \mathbb{R}$ is an investment target set at $t = 0$ as the desired mean payoff at the end of the investment horizon $[0, T]$.

Due to the variance in its objective, (3) is known to be time inconsistent. The problem then becomes descriptive rather than normative because there is generally no dynamically optimal solution for a time-inconsistent optimization problem. Agents react differently to a same time-inconsistent problem, and a goal of the study becomes to describe the different behaviors when facing such time-inconsistency. In this paper we focus ourselves to the so-called pre-committed strategies of the MV problem, which are optimal at $t = 0$ only.

---

1In practice, the true (yet unknown) investment opportunity parameters $\mu$, $\sigma$ and $r$ can be time-varying stochastic processes. Most existing quantitative finance methods are devoted to estimating these parameters. In contrast, RL learns the values of various strategies and the optimal value through exploration and exploitation, without assuming any statistical properties of these parameters or estimating them. But for a model-based, classical MV problem, we assume these parameters are constant and known. In subsequent contexts, all we need is the structure of the problem for our RL algorithm design.

2The original MV problem is to find the Pareto efficient frontier for a two-objective (i.e. maximizing the expected terminal payoff and minimizing its variance) optimization problem. There are a number of equivalent mathematical formulations to finding such a frontier, (3) being one of them. In particular, by varying the parameter $z$ one can trace out the frontier. See Zhou and Li (2000) for details.

3For a detailed discussions about the different behaviors under time-inconsistency, see
To solve (3), one first transforms it into an unconstrained problem by applying a Lagrange multiplier \( w \):

\[
\min_u E[(x_u^T)^2] - z^2 - 2w (E[x_U] - z) = \min_u E[(x_u^T - w)^2] - (w - z)^2. \tag{4}
\]

This problem can be solved analytically, whose solution \( u^* = \{u^*_t, 0 \leq t \leq T\} \) depends on \( w \). Then the original constraint \( E[x_u^T] = z \) determines the value of \( w \). We refer a detailed derivation to Zhou and Li (2000).

### 2.2 Exploratory continuous-time MV problem

Given the complete knowledge of the model parameters, the classical, model-based MV problem (3) and many of its variants have been solved rather completely. When implementing these solutions, one needs to estimate the market parameters from historical time series of asset prices, a procedure known as **identification** in classical adaptive control. However, it is well known that in practice it is difficult to estimate the investment opportunity parameters, especially the mean return (aka the **mean–blur problem**; see, e.g., Luenberger (1998)) with a workable accuracy. Moreover, the classical optimal MV strategies are often extremely sensitive to these parameters, largely due to the procedure of inverting ill-conditioned variance–covariance matrices to obtain optimal allocation weights.

On the other hand, RL techniques do not require, and indeed often skip, any estimation of model parameters. Rather, RL algorithms, driven by historical data, output optimal (or near-optimal) allocations directly. This is made possible by direct interactions with the unknown investment environment, in a learning (exploring) while optimizing (exploiting) fashion. Wang et al. (2019) motivate and propose a theoretical framework for general exploratory, RL stochastic control and carry out a detailed study for the special linear–quadratic (LQ) case, albeit in the setting of the infinite time horizon. We adopt the same framework here, while noting the inherent features of an LQ structure and a **finite** time horizon of the MV problem. Indeed, although the motivation for the exploratory formulation is mostly the same, there are intriguing new insights emerging with this transition from the infinite time horizon to the finite counterpart.

First, we introduce the “exploratory” version of the state dynamics (2). It was originally proposed in Wang et al. (2013), motivated by repet-

---

\(^4^\)Strictly speaking, \( 2w \in \mathbb{R} \) is the Lagrange multiplier.
itive learning in RL. In this formulation, the control (portfolio) process
\( u = \{u_t, 0 \leq t \leq T\} \) is randomized, which represents exploration and learning, leading to a measure-valued or distributional control process whose density function is given by \( \pi = \{\pi_t, 0 \leq t \leq T\} \). The dynamics \( \mathcal{P}(\mathbb{R}) \) is changed to
\[
\begin{align*}
    dX^\pi_t &= \tilde{b}(\pi_t) \, dt + \tilde{\sigma}(\pi_t) \, dW_t, \\
    \tilde{b}(\pi) &= \int_{\mathbb{R}} \rho u \pi(u) \, du, \quad \pi \in \mathcal{P}(\mathbb{R}), \\
    \tilde{\sigma}(\pi) &= \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi(u) \, du}, \quad \pi \in \mathcal{P}(\mathbb{R}),
\end{align*}
\]
where \( 0 < t \leq T \) and \( X^\pi_0 = x \). The randomized, distributional control process \( \pi = \{\pi_t, t \in [0, T]\} \) is to model exploration, whose overall level is in turn captured by its accumulative differential entropy
\[
    \mathcal{H}(\pi) := -\int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) \, du \, dt.
\]
Further, introduce a \textit{temperature parameter} (or \textit{exploration weight}) \( \lambda > 0 \) reflecting the trade-off between exploitation and exploration. The entropy-regularized, exploratory MV problem is then to solve, for any fixed \( w \in \mathbb{R} \):
\[
    \inf_{\pi \in \mathcal{A}(x,0)} \mathbb{E} \left[ (X^\pi_T - w)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) \, du \, dt | X^\pi_0 = x \right] = (w - z)^2,
\]
where $\mathcal{A}(x,0)$ is the set of admissible distributional controls on $[0,T]$ to be precisely defined below. Once this problem is solved with a minimizer $\pi^* = \{\pi^*_t, 0 \leq t \leq T\}$, the Lagrange multiplier $w$ can be determined by the additional constraint $E[X^\pi_T] = z$.

The optimization objective (11) explicitly encourages exploration, in contrast to the classical problem (4) which concerns exploitation only.

We will solve (11) by dynamic programming. For that we need to define the value functions. For each $(s,y) \in [0,T) \times \mathbb{R}$, consider the state equation (9) on $[s,T]$ with $X^\pi_s = y$. Define the set of admissible controls, $\mathcal{A}(s,y)$, as follows. Let $\mathcal{B}(\mathbb{R})$ be the Borel algebra on $\mathbb{R}$. A (distributional) control (or portfolio/strategy) process $\pi = \{\pi_t, s \leq t \leq T\}$ belongs to $\mathcal{A}(s,y)$, if

(i) for each $s \leq t \leq T$, $\pi_t \in \mathcal{P}(U)$ a.s.;

(ii) for each $A \in \mathcal{B}(\mathbb{R})$, $\int_A \pi_t(u)du, s \leq t \leq T$ is $\mathcal{F}_t$-progressively measurable;

(iii) $E\left[\int_s^T (\mu_t^2 + \sigma_t^2) dt\right] < \infty$;

(iv) $E\left[(X^\pi_T - y)^2 + \lambda \int_s^T \int_\mathbb{R} \pi_t(u) \ln \pi_t(u)dudt\right] < \infty$.

Clearly, it follows from condition (iii) that the stochastic differential equation (SDE) (9) has a unique strong solution for $s \leq t \leq T$ that satisfies $X^\pi_s = y$.

Controls in $\mathcal{A}(s,y)$ are measure-valued (or, more precisely, density-function-valued) stochastic processes, which are also called open-loop controls in the control terminology. As in classical control theory, it is important to distinguish between open-loop controls and feedback (or closed-loop) controls (or policies or laws). Specifically, a deterministic mapping $\pi(\cdot; t, x)$ is called an (admissible) feedback control if i) $\pi(\cdot; t, x)$ is a density function for each $(t, x) \in [0,T] \times \mathbb{R}$; ii) for each $(s,y) \in [0,T) \times \mathbb{R}$, the following SDE (which is the system dynamics after the feedback policy $\pi(\cdot; t, x)$ is applied)

$$dX^\pi_t = \tilde{b}(\pi(\cdot; t, X^\pi_t))dt + \tilde{\sigma}(\pi(\cdot; t, X^\pi_t))dW_t, \ t \in [s,T]; \ X^\pi_s = y$$ (12)

has a unique strong solution $\{X^\pi_t, t \in [s,T]\}$, and the open-loop control $\pi = \{\pi_t, t \in [s,T]\} \in \mathcal{A}(s,y)$ where $\pi_t := \pi(\cdot; t, X^\pi_t)$. In this case, the open-loop control $\pi$ is said to be generated from the feedback policy $\pi(\cdot; t, x)$ with respect to the initial time and state, $(s,y)$. It is useful to note that an open-loop control and its admissibility depend on the initial $(s,y)$, whereas $^5$Actually (9) is a trivial SDE in that its right hand side does not depend on the unknown $X^\pi$. 

9
a feedback policy can generate open-loop controls for any \((s, y) \in [0, T) \times \mathbb{R}\), and hence is in itself independent of \((s, y)\).

Now, for a fixed \(w \in \mathbb{R}\), define

\[
V(s, y; w) := \inf_{\pi \in \mathcal{A}(s, y)} \mathbb{E} \left[ (X_T^\pi - w)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du dv \mid X_s^\pi = y \right] - (w - z)^2,
\]

for \((s, y) \in [0, T) \times \mathbb{R}\). The function \(V(\cdot, \cdot; w)\) is called the optimal value function of the problem. Moreover, we define the value function under any given feedback control \(\pi\):

\[
V^\pi(s, y; w) = \mathbb{E} \left[ (X_T^\pi - w)^2 + \lambda \int_s^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du dv \mid X_s^\pi = y \right] - (w - z)^2,
\]

for \((s, y) \in [0, T) \times \mathbb{R}\), where \(\pi = \{\pi_t, t \in [s, T]\}\) is the open-loop control generated from \(\pi\) with respect to \((s, y)\) and \(\{X_t^\pi, t \in [s, T]\}\) is the corresponding wealth process.

### 3 Solving Exploratory MV Problem

In this section we first solve the exploratory MV problem, and then establish solvability equivalence between the classical and exploratory problems. The latter is important for understanding the cost of exploration and for devising RL algorithms.

#### 3.1 Optimality of Gaussian exploration

To solve the exploratory MV problem, we apply the classical Bellman’s principle of optimality:

\[
V(t, x; w) = \inf_{\pi \in \mathcal{A}(t, x)} \mathbb{E} \left[ V(s, X_s^\pi; w) + \lambda \int_t^s \int_{\mathbb{R}} \pi_v(u) \ln \pi_v(u) du dv \mid X_t^\pi = x \right],
\]

---

6 Throughout this paper, we use boldfaced \(\pi\) to denote feedback controls, and the normal style \(\pi\) to denote open-loop controls.

7 In the control literature, \(V\) is called the value function. However, in the RL literature the term “value function” is also used for the objective value under a particular control. So to avoid ambiguity we call \(V\) the optimal value function.
for $x \in \mathbb{R}$ and $0 \leq t < s \leq T$. Following standard arguments, we deduce that $V$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$
v_t(t, x; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \left( \frac{1}{2} \sigma^2 \pi v_{xx}(t, x; w) + \tilde{b}(x, \pi) v_x(t, x; w) + \lambda \int_{\mathbb{R}} \pi(u) \ln \pi(u) du \right) = 0,
$$

or, equivalently,

$$
v_t(t, x; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 v_{xx}(t, x; w) + \rho \sigma u v_x(t, x; w) + \lambda \ln \pi(u) \right) \pi(u) du = 0,
$$

with the terminal condition $v(T, x; w) = (x-w)^2 - (w-z)^2$. Here $v$ denotes the generic unknown solution to the HJB equation.

Applying the usual verification technique and using the fact that $\pi \in \mathcal{P}(\mathbb{R})$ if and only if

$$
\int_{\mathbb{R}} \pi(u) du = 1 \quad \text{and} \quad \pi(u) \geq 0 \ \text{a.e. on} \ \mathbb{R},
$$

we can solve the (constrained) optimization problem in the HJB equation (16) to obtain a feedback (distributional) control whose density function is given by

$$
\pi^*(u; t, x, w) = \frac{\exp \left( -\frac{1}{2} \left( \frac{1}{2} \sigma^2 u^2 v_{xx}(t, x; w) + \rho \sigma u v_x(t, x; w) \right) \right)}{\int_{\mathbb{R}} \exp \left( -\frac{1}{2} \left( \frac{1}{2} \sigma^2 u^2 v_{xx}(t, x; w) + \rho \sigma u v_x(t, x; w) \right) \right) du},
$$

where we have denoted by $\mathcal{N}(u|\alpha, \beta)$ the Gaussian density function with mean $\alpha \in \mathbb{R}$ and variance $\beta > 0$. In the above representation, we have assumed that $v_{xx}(t, x; w) > 0$, which will be verified in what follows.

Substituting the candidate optimal Gaussian feedback control policy (18) back into the HJB equation (16), the latter is transformed to

$$
v_t(t, x; w) - \frac{\rho^2}{2} v_x^2(t, x; w) + \frac{\lambda}{2} \left( 1 - \ln \frac{2\pi e \lambda}{\sigma^2 v_{xx}(t, x; w)} \right) = 0,
$$

with $v(T, x; w) = (x-w)^2 - (w-z)^2$. A direct computation yields that this equation has a classical solution 

$$
v(t, x; w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda \rho^2}{4} \left( T^2 - t^2 \right) - \frac{\lambda}{2} \left( \rho^2 T - \ln \frac{\sigma^2}{\pi \lambda} \right) (T-t)-(w-z)^2,
$$

(20)
which clearly satisfies $v_{xx}(t, x; w) > 0$, for any $(t, x) \in [0, T] \times \mathbb{R}$. It then follows that the candidate optimal feedback Gaussian control (18) reduces to

$$\pi^*(u; t, x, w) = \mathcal{N}(u \mid -\frac{\rho}{\sigma}(x - w), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}), \quad (t, x) \in [0, T] \times \mathbb{R}. \quad (21)$$

Finally, the optimal wealth process (9) under $\pi^*$ becomes

$$dX^*_t = -\rho^2(X^*_t - w) dt + \sqrt{\rho^2 (X^*_t - w)^2 + \frac{\lambda}{2} e^{\rho^2(T-t)}} dW_t, \quad X^*_0 = x. \quad (22)$$

It has a unique strong solution for $0 \leq t \leq T$, as can be easily verified.

We now summarize the above results in the following theorem.

**Theorem 1** The value function of the entropy-regularized exploratory MV problem (11) is given by

$$V(t, x; w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda \rho^2}{4} (T^2 - t^2) - \frac{\lambda}{2} \left( \rho^2 T - \ln \frac{\sigma^2}{\pi \lambda} \right) (T-t) - (w-z)^2, \quad (23)$$

for $(t, x) \in [0, T] \times \mathbb{R}$. Moreover, the optimal feedback control is Gaussian, with its density function given by

$$\pi^*(u; t, x, w) = \mathcal{N}(u \mid -\frac{\rho}{\sigma}(x - w), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}). \quad (24)$$

The associated optimal wealth process under $\pi^*$ is the unique solution of the SDE

$$dX^*_t = -\rho^2(X^*_t - w) dt + \sqrt{\rho^2 (X^*_t - w)^2 + \frac{\lambda}{2} e^{\rho^2(T-t)}} dW_t, \quad X^*_0 = x. \quad (25)$$

Finally, the Lagrange multiplier $w$ is given by $w = \frac{ze^{\rho^2(T-t)-x}}{e^{\rho^2(T-t)-1}}$.

**Proof.** For each fixed $w \in \mathbb{R}$, the verification arguments aim to show that the value function of problem (11) is given by (23) and that the candidate optimal policy (24) is indeed admissible. A detailed proof follows the same lines of that of Theorem 4 in [Wang et al., 2019], and is left for interested readers.

We now determine the Lagrange multiplier $w$ through the constraint $\mathbb{E}[X^*_T] = z$. It follows from (25), along with the standard estimate that $\mathbb{E}[\max_{t \in [0, T]} (X^*_t)^2] < \infty$ and Fubini’s Theorem, that

$$\mathbb{E}[X^*_t] = x + \mathbb{E} \left[ \int_0^t -\rho^2(X^*_s - w) ds \right] = x + \int_0^t -\rho^2 (\mathbb{E}[X^*_s] - w) ds.$$
Hence, $E[X^*_t] = (x - w)e^{-\rho^2 t} + w$. The constraint $E[X^*_T] = z$ now becomes $(x - w)e^{-\rho^2 T} + w = z$, which gives $w = \frac{xe^{\rho^2 T} - z}{e^{\rho^2 T} - 1}$.

There are several interesting points to note in this result. First of all, it follows from Theorem 2 in the next section that the classical and the exploratory MV problems have the same Lagrange multiplier value due to the fact that the optimal terminal wealths under the respective optimal feedback controls of the two problems turn out to have the same mean. This latter result is rather surprising at first sight because the exploration greatly alters the underlying system dynamics (compare the dynamics (2) with (5)).

Second, the variance of the optimal Gaussian control, which measures the level of exploration, is $\frac{\lambda^2 \sigma^2}{2\rho^2} e^{\rho^2 (T-t)}$ at time $t$. So the exploration decays in time: the agent initially engages in exploration at the maximum level, and reduces it gradually (although never to zero) as time passes and approaches the end of the investment horizon. Hence, different from its infinite horizon counterpart studied in [Wang et al. (2019)], the extent of exploration is no longer constant, but, rather, annealing. This is intuitive because, as she learns more about the random environment as time passes, the exploitation becomes more important since there is a deadline $T$ at which her actions will be evaluated. Naturally, exploitation dominates exploration as time nears maturity. Theorem 1 presents such a decaying exploration scheme endogenously which, to our best knowledge, has not been derived in the RL literature.

Third, as already noted in [Wang et al. (2019)], at any given $t \in [0, T]$, the variance of the exploratory Gaussian distribution decreases as the volatility of the risky asset increases, with other parameters being fixed. The volatility of the risky asset reflects the level of randomness of the investment universe. This hints that a more random environment contains more learning opportunities, which the agent can leverage to reduce her own exploratory endeavor because, after all, exploration is costly.

Finally, the mean of the Gaussian distribution (24) is independent of the exploration weight $\lambda$, while its variance is independent of the state $x$. This highlights a perfect separation between exploitation and exploration, as the former is captured by the mean and the latter by the variance of the optimal Gaussian exploration. This property is also consistent with the case of infinite horizon studied in [Wang et al. (2019)].

---

8Theorem 2 is a reproduction of the results on the classical MV problem obtained in Zhou and Li (2000).

13
3.2 Solvability equivalence between classical and exploratory MV problems

In this section, we establish the solvability equivalence between the classical and the exploratory, entropy-regularized MV problems. Note that both problems can be and indeed have been solved separately and independently. Here by “solvability equivalence” we mean that the solution of one problem will lead to that of the other directly, without needing to solve it separately. This equivalence was discovered in [Wang et al. (2019)] for the infinite horizon case, and was shown to be instrumental in deriving the convergence result (when the exploration weight $\lambda$ decays to 0) as well as analyzing the exploration cost therein. Here, the discussions are mostly parallel; so we will be brief.

Recall the classical MV problem (4). In order to apply dynamic programming, we again consider the set of admissible controls, $A^{cl}(s, y)$, for $(s, y) \in [0, T) \times \mathbb{R}$:

$$A^{cl}(s, y) := \left\{ u = \{u_t, t \in [s, T] \} : u \text{ is } \mathcal{F}_t\text{-progressively measurable and } \mathbb{E} \left[ \int_s^T (u_s)^2 \, ds \right] < \infty \right\}.$$  

The (optimal) value function is defined by

$$V^{cl}(s, y; w) = \inf_{u \in A^{cl}(s,y)} \mathbb{E} \left[ (x^u_T - w)^2 \mid x^u_s = y \right] - (w - z)^2, \quad (26)$$

for $(s, y) \in [0, T) \times \mathbb{R}$, where $w \in \mathbb{R}$ is fixed. Once this problem is solved, $w$ can be determined by the constraint $\mathbb{E}[x^*_T] = z$, with $\{x^*_t, t \in [0, T]\}$ being the optimal wealth process under the optimal portfolio $u^*$.

The HJB equation is

$$\omega(t, x; w) + \min_{u \in \mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 \omega_{xx}(t, x; w) + \rho \sigma u \omega_x(t, x; w) \right) = 0, \quad (t, x) \in [0, T) \times \mathbb{R},$$

with the terminal condition $\omega(t, x; w) = (x - w)^2 - (w - z)^2$.

Standard verification arguments deduce the value function to be

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

the optimal feedback control policy to be

$$u^*(u; t, x, w) = -\frac{\rho}{\sigma} (x - w), \quad (28)$$
and the corresponding optimal wealth process to be the unique strong solution to the SDE
\[ dx^*_t = -\rho^2 (x^*_t - w) \, dt - \rho (x^*_t - w) \, dW_t, \quad x^*_0 = x. \] (29)
Comparing the optimal wealth dynamics, (25) and (29), of the exploratory and classical problems, we note that they have the same drift coefficient (but different diffusion coefficients). As a result, the two problems have the same mean of terminal optimal wealth and hence the same value of the Lagrange multiplier \( w = ze^{\rho^2 T - x} \) determined by the constraint \( E[x^*_T] = z \).

We now provide the solvability equivalence between the two problems. The proof is very similar to that of Theorem 7 in Wang et al. (2019), and is thus omitted.

**Theorem 2** The following two statements (a) and (b) are equivalent.

(a) The function \( v(t, x; w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda^2}{4} (T^2 - t^2) - \frac{\lambda}{2} \left( \rho^2 T - \ln \frac{e^{2\rho^2 T - x}}{\pi \lambda} \right) (T-t) - (w-z)^2, \) \( (t, x) \in [0, T] \times \mathbb{R}, \) is the value function of the exploratory MV problem (11), and the corresponding optimal feedback control is
\[ \pi^*(u; t, x, w) = N \left( u \bigg| -\frac{\rho}{\sigma} (x - w), \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)} \right). \]

(b) The function \( \omega(t, x; w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2, \) \( (t, x) \in [0, T] \times \mathbb{R}, \) is the value function of the classical MV problem (26), and the corresponding optimal feedback control is
\[ u^*(t, x; w) = -\frac{\rho}{\sigma} (x - w). \]

Moreover, the two problems have the same Lagrange multiplier \( w = ze^{\rho^2 T - x} \).

It is reasonable to expect that the exploratory problem converges to its classical counterpart as the exploration weight \( \lambda \) decreases to 0. The following result makes this precise.

**Theorem 3** Assume that statement (a) (or equivalently, (b)) of Theorem 2 holds. Then, for each \( (t, x, w) \in [0, T] \times \mathbb{R} \times \mathbb{R}, \)
\[ \lim_{\lambda \to 0} \pi^*_\lambda(\cdot; t, x; w) = \delta_{u^*(t, x; w)}(\cdot) \] weakly.
Moreover,
\[ \lim_{\lambda \to 0} |V(t, x; w) - V^{cl}(t, x; w)| = 0. \]
Proof. The weak convergence of the feedback controls follows from the explicit forms of $\pi^*$ and $u^*$ in statements (a) and (b). The pointwise convergence of the value functions follows easily from the forms of $V(\cdot)$ and $V^{\text{cl}}(\cdot)$, together with the fact that
\[
\lim_{\lambda \to 0} \frac{\lambda}{2} \ln \frac{\sigma^2}{\pi \lambda} = 0.
\]

Finally, we conclude this section by examining the cost of the exploration. This was originally defined and derived in Wang et al. (2019) for the infinite horizon setting. Here, the cost associated with the MV problem due to the explicit inclusion of exploration in the objective (11) is defined by

\[
C^{u^*, \pi^*}(0, x; w) := \left( V(0, x; w) - \lambda \mathbb{E} \left[ \int_0^T \int_{\mathbb{R}} \pi^*_t(u) \ln \pi^*_t(u) du \, dt \bigg| X_{0}^{\pi^*} = x \right] \right)
- V^{\text{cl}}(0, x; w),
\]

for $x \in \mathbb{R}$, where $\pi^* = \{ \pi^*_t, t \in [0, T] \}$ is the (open-loop) optimal strategy generated by the optimal feedback law $\pi^*$ with respect to the initial condition $X_{0}^{\pi^*} = x$. This cost is the difference between the two optimal value functions, adjusting for the additional contribution due to the entropy value of the optimal exploratory strategy.

Making use of Theorem 2, we have the following result.

**Theorem 4** Assume that statement (a) (or equivalently, (b)) of Theorem 2 holds. Then, the exploration cost for the MV problem is
\[
C^{u^*, \pi^*}(0, x; w) = \frac{\lambda T}{2}, \quad x \in \mathbb{R}, \ w \in \mathbb{R}.
\]

Proof. Let $\{\pi^*_t, t \geq 0\}$ be the open-loop control generated by the feedback control $\pi^*$ given in statement (a) with respect to the initial state $x$ at $t = 0$, namely,
\[
\pi^*_t(u) = \mathcal{N} \left( u \bigg| - \frac{\rho}{\sigma^2} (X^*_t - w), \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)} \right)
\]
where $\{X^*_t, t \in [0, T]\}$ is the corresponding optimal wealth process of the exploratory MV problem, starting from the state $x$ at $t = 0$, when $\pi^*$ is applied. Then, we easily deduce that
\[
\int_{\mathbb{R}} \pi^*_t(u) \ln \pi^*_t(u) du = -\frac{1}{2} \ln \left( \frac{\pi e \lambda}{\sigma^2} e^{\rho^2(T-t)} \right).
\]
The desired result now follows immediately from the expressions of $V(\cdot)$ in (a) and $V_{\text{cl}}(\cdot)$ in (b).

The exploration cost depends on only two “agent-specific” parameters, the exploration weight $\lambda > 0$ and the investment horizon $T > 0$. Note that the latter is simultaneously the exploration horizon. Our result is intuitive in that the exploration cost increases with the exploration weight and with the exploration horizon. Indeed, the dependence is linear with respect to each of the two attributes: $\lambda$ and $T$. It is also interesting to note that the cost is independent of the Lagrange multiplier. This 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 $w$.

4 RL Algorithm Design

Having laid the theoretical foundation in the previous two sections, we now design an RL algorithm to learn the solution of the entropy-regularized MV problem and to output implementable portfolio allocation strategies, without assuming any knowledge about the underlying parameters. To this end, we will first establish a so-called policy improvement theorem as well as the corresponding convergence result. Meanwhile, we will also provide a self-correcting scheme to learn the true Lagrange multiplier $w$, based on stochastic approximation. Our RL algorithm bypasses the phase of estimating any model parameters, including the mean return vector and the variance-covariance matrix. It also avoids inverting a typically ill-conditioned variance-covariance matrix in high dimensions that would likely produce non-robust allocations.

In this paper, rather than relying on the typical framework of discrete-time Markov Decision Processes (MDP) for most RL problems and discretizing time and space accordingly, we design an algorithm to learn the solutions of the continuous-time exploratory MV problem (11) directly. Specifically, we adopt the method developed in Doya (2000) to avoid discretization of the state dynamics or the HJB equation. As pointed out in Doya (2000), it is typically challenging to find the right granularity to discretize the state, action and time, and naive discretization may lead to poor perfor-

---

9In Wang et al. (2019) with the infinite horizon, an analogous result is obtained which states that exploration cost is proportional to the exploration weight and inversely proportional to the discount factor. Clearly, here the length of time horizon, $T$, plays a role similar to the inverse of the discount factor.
formance. On the other hand, grid-based discretization methods for solving the HJB equation cannot be easily extended to high-dimensional state space in practice due to the curse of dimensionality, although theoretical convergence results have been established (see Munos and Bourgine (1998); Munos (2000)). Our algorithm, to be described in the next section, however, makes use of the provable policy improvement theorem and fairly simple yet efficient functional approximations to directly learn the value functions and the optimal Gaussian policy. Moreover, it is computationally feasible and implementable in high-dimensional state spaces (i.e., with a large number of risky assets) due to the explicit representation for the value functions and the portfolio strategies, thereby devoid of the curse of dimensionality. Note that our algorithm does not use (deep) neural networks, which have been applied extensively in literature for (high-dimensional) continuous RL problems (e.g., Lillicrap et al. (2016), Mnih et al. (2015)), but known for unstable performance, sample inefficiency as well as extensive hyperparameter tuning (Mnih et al. (2015)), in addition to their low interpretability.

4.1 A policy improvement theorem

Most RL algorithms consist of two iterative procedures: policy evaluation and policy improvement (Sutton and Barto (2018)). The former provides an estimated value function for the current policy, whereas the latter updates the current policy in the right direction to improve the value function. A policy improvement theorem (PIT) is therefore a crucial prerequisite for interpretable RL algorithms that ensures the iterated value functions to be non-increasing (in the case of a minimization problem), and ultimately converge to the optimal value function; see, for example, Section 4.2 of Sutton and Barto (2018). PITs have been proved for discrete-time entropy-regularized RL problems in infinite horizon (Haarnoja et al. (2017)), and for continuous-time classical stochastic control problems (Jacka and Mijatović (2017)). The following result provides a PIT for our exploratory MV portfolio selection problem.

Theorem 5 (Policy Improvement Theorem) Let $w \in \mathbb{R}$ be fixed and $\pi = \pi(\cdot; \cdot; \cdot, w)$ be an admissible feedback control policy. Suppose that the

\footnote{Interpretability is one of the most important and pressing issues in the general artificial intelligence applications in financial industry due to, among others, the regulatory requirement.}

\footnote{Jacka and Mijatović (2017) consider classical stochastic control problems with no distributional controls nor entropy regularization. Their focus was not on RL and related issues including exploration.}
corresponding value function $V^\pi(\cdot;v,w) \in C^{1,2}([0,T] \times \mathbb{R}) \cap C^0([0,T] \times \mathbb{R})$ and satisfies $V^\pi_{xx}(t,x;w) > 0$, for any $(t,x) \in [0,T] \times \mathbb{R}$. Suppose further that the feedback policy $\tilde{\pi}$ defined by

$$\tilde{\pi}(u; t, x, w) = \mathcal{N} \left( u \left| - \frac{\rho V^\pi_{xx}(t,x;w)}{\sigma^2 V^\pi_{xx}(t,x;w)} \right., \left. \frac{\lambda}{\sigma^2 V^\pi_{xx}(t,x;w)} \right) \right)$$

is admissible. Then,

$$V^\pi(t, x; w) \leq V^\pi(t, x; w), \quad (t,x) \in [0,T] \times \mathbb{R}. \quad (33)$$

**Proof.** Fix $(t, x) \in [0, T] \times \mathbb{R}$. Since, by assumption, the feedback policy $\tilde{\pi}$ is admissible, the open-loop control strategy, $\pi = \{\bar{\pi}_v, v \in [t,T]\}$, generated from $\tilde{\pi}$ with respect to the initial condition $X^\pi_t = x$ is admissible. Let $\{X^\pi_t, s \in [t,T]\}$ be the corresponding wealth process under $\tilde{\pi}$. Applying Itô’s formula, we have

$$V^\pi(s, \bar{X}_s) = V^\pi(t, x) + \int_t^s V^\pi_t(v, X^\pi_v) dv + \int_t^s \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 V^\pi_{xx}(v, X^\pi_v) \right) \bar{\pi}_v(u) du dv$$

$$+ \rho \sigma u V^\pi_x(v, X^\pi_v) \bar{\pi}_v(u) du + \int_t^s \sigma \left( \int_{\mathbb{R}} u^2 \bar{\pi}_v(u) du \right)^{\frac{1}{2}} V^\pi_x(v, X^\pi_v) dW_v, \quad s \in [t,T].$$

Define the stopping times $\tau_n := \inf \{s \geq t : \int_t^s \sigma^2 \int_{\mathbb{R}} u^2 \bar{\pi}_v(u) du (V^\pi_x(v, X^\pi_v))^2 dv \geq n\}, \quad n \geq 1$. Then, from (34), we obtain

$$V^\pi(t, x) = \mathbb{E} \left[ V^\pi(s \wedge \tau_n, X^\pi_{s \wedge \tau_n}) - \int_t^{s \wedge \tau_n} V^\pi_t(v, X^\pi_v) dv \right.$$

$$- \int_t^{s \wedge \tau_n} \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 V^\pi_{xx}(v, X^\pi_v) + \rho \sigma u V^\pi_x(v, X^\pi_v) \right) \bar{\pi}_v(u) du dv \left| X^\pi_t = x \right].$$

On the other hand, by standard arguments and the assumption that $V^\pi$ is smooth, we have

$$V^\pi_t(t, x) + \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 V^\pi_{xx}(t, x) + \rho \sigma u V^\pi_x(t, x) + \lambda \ln \pi(u; t, x) \right) \pi(u; t, x) du = 0,$$

for any $(t, x) \in [0, T] \times \mathbb{R}$. It follows that

$$V^\pi_t(t, x) + \min_{\pi' \in P(\mathbb{R})} \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 V^\pi_{xx}(t, x) + \rho \sigma u V^\pi_x(t, x) + \lambda \ln \pi'(u) \right) \pi'(u) du \leq 0.$$  

(36)
Notice that the minimizer of the Hamiltonian in (36) is given by the feedback policy \( \tilde{\pi} \) in (32). It then follows that equation (35) implies

\[
V_\pi(t, x) \geq E \left[ V_\pi(s \wedge \tau_n, X_{s \wedge \tau_n}) + \lambda \int_t^{s \wedge \tau_n} \pi_v(u) \ln \pi_v(u) \, du \bigg| X_t = x \right],
\]

for \((t, x) \in [0, T] \times \mathbb{R}\) and \(s \in [t, T]\). Now taking \(s = T\), and using that \(V_\pi(T, x) = V_{\tilde{\pi}}(T, x) = (x-w)^2 - (w-z)^2\) together with the assumption that \(\tilde{\pi}\) is admissible, we obtain, by sending \(n \to \infty\) and applying the dominated convergence theorem, that

\[
V_\pi(t, x) \geq E \left[ V_{\tilde{\pi}}(T, X_{T}) + \lambda \int_t^T \pi_v(u) \ln \pi_v(u) \, du \bigg| X_t = x \right] = V_{\tilde{\pi}}(t, x),
\]

for any \((t, x) \in [0, T] \times \mathbb{R}\).

The above policy improvement theorem suggests that there are always policies in the Gaussian family to improve the value function, regardless of what an initial policy the RL agent may start with. Hence, without loss of generality, we can simply focus on the Gaussian policies when choosing an initial solution. Moreover, the optimal Gaussian policy (24) in Theorem 1 suggests that a candidate initial feedback policy may take the form

\[
\pi_0(u; t, x, w) = \mathcal{N}(u|a(x-w), c_1 e^{c_2(T-t)})
\]

It turns out that, even theoretically, such a choice leads to the convergence of both the value functions and the policies in a finite number of iterations.

**Theorem 6** Let \(\pi_0(u; t, x, w) = \mathcal{N}(u|a(x-w), c_1 e^{c_2(T-t)})\), with \(a, c_2 \in \mathbb{R}\) and \(c_1 > 0\). Denote by \(\{\pi_n(u; t, x, w), (t, x) \in [0, T] \times \mathbb{R}, n \geq 1\}\) the sequence of feedback policies updated by the policy improvement scheme (32), and \(\{V^{\pi_n}(t, x, w), (t, x) \in [0, T] \times \mathbb{R}, n \geq 1\}\) the sequence of the corresponding value functions. Then,

\[
\lim_{n \to \infty} \pi_n(\cdot; t, x, w) = \pi^*(\cdot; t, x, w) \quad \text{weakly}, \tag{37}
\]

and

\[
\lim_{n \to \infty} V^{\pi_n}(t, x, w) = V(t, x, w), \tag{38}
\]

for any \((t, x, w) \in [0, T] \times \mathbb{R} \times \mathbb{R}\), where \(\pi^*\) and \(V\) are the optimal Gaussian policy (24) and the optimal value function (23), respectively.

**Proof.** It can be easily verified that the feedback policy \(\pi_0\) where \(\pi_0(u; t, x, w) = \mathcal{N}(u|a(x-w), c_1 e^{c_2(T-t)})\) generates an open-loop policy \(\tilde{\pi}_0\) that is admissible
with respect to the initial \((t, x)\). Moreover, it follows from the Feynman-Kac formula that the corresponding value function \(V^{\pi_0}\) satisfies the PDE

\[
V_t^{\pi_0}(t, x; w) + \int_{\mathbb{R}} \left( \frac{1}{2} \sigma^2 u^2 V_{xx}^{\pi_0}(t, x; w) + \rho \sigma u V_x^{\pi_0}(t, x; w) + \lambda \ln \pi_0(u; t, x, w) \right) \pi_0(u; t, x, w) du = 0,
\]

with terminal condition \(V^{\pi_0}(T, x; w) = (x - w)^2 - (w - z)^2\). Solving this equation we obtain

\[
V^{\pi_0}(t, x; w) = (x - w)^2 e^{(2\rho \sigma a + \sigma^2 a^2)(T - t)} + \int_t^T c_1 \sigma^2 e^{(2\rho \sigma a + \sigma^2 a^2 + c_2)(T - s)} ds + \frac{\lambda c_2}{4} (T - t)^2 + \frac{\lambda \ln(2\pi e c_1)}{2}(T - t) - (w - z)^2.
\]

It is easy to check that \(V^{\pi_0}\) satisfies the conditions in Theorem 5 and, hence, the theorem applies. The improved policy is given by (32), which, in the current case, becomes

\[
\pi_1(u; t, x, w) = \mathcal{N} \left( u \mid -\frac{\rho}{\sigma} (x - w), \frac{\lambda}{2\sigma^2 e^{(2\rho \sigma a + \sigma^2 a^2)(T - t)}} \right).
\]

Again, we can calculate the corresponding value function as \(V^{\pi_1}(t, x; w) = (x - w)^2 e^{-\rho^2(T - t)} + F_1(t)\), where \(F_1\) is a function of \(t\) only. Theorem 5 is applicable again, which yields the improved policy \(\pi_2\) as exactly the optimal Gaussian policy \(\pi^*\) given in (24), together with the optimal value function \(V\) in (23). The convergence therefore follows, as for \(n \geq 2\), both the policy and the value function will no longer strictly improve under the policy improvement scheme (32).

The above convergence result shows that if we choose the initial policy wisely, the learning scheme will, theoretically, converge after a finite number of (two, in fact) iterations. When implementing this scheme in practice, of course, the value function for each policy can only be approximated and, hence, the learning process typically takes more iterations to converge. Nevertheless, Theorem 5 provides the theoretical foundation for updating a current policy, while Theorem 6 suggests a good starting point in the policy space. We will make use of both results in the next subsection to design an implementable RL algorithm for the exploratory MV problem.
4.2 The ENT-MV algorithm

In this section, we present the (continuous-time) entropy-regularized Mean–variance RL algorithm, the ENT-MV algorithm, to solve (11). It consists of three concurrently ongoing procedures: the policy evaluation, the policy improvement and a self-correcting scheme for learning the Lagrange multiplier \( w \) based on stochastic approximation.

For the policy evaluation, we follow the method employed in Doya (2000) for learning the value function \( V^\pi \) under any arbitrarily given admissible feedback policy \( \pi \). By Bellman’s consistency, we have

\[
V^\pi(t, x) = E \left[ V^\pi(X_s, s) + \lambda \int_t^s \int_R \pi_v(u) \ln \pi_v(u) dudv | X_t = x \right], \quad s \in [t, T],
\]

for \((t, x) \in [0, T] \times \mathbb{R}\). Rearranging this equation and dividing both sides by \( s - t \), we obtain

\[
E \left[ \frac{V^\pi(X_s, s) - V^\pi(X_t, t)}{s - t} \right] + \lambda \int_t^s \int_R \pi_v(u) \ln \pi_v(u) dudv | X_t = x = 0.
\]

Taking \( s \to t \) gives rise to the continuous-time Bellman’s error (or the temporal difference (TD) error; see Doya (2000)):

\[
\delta_t := \dot{V}_t^\pi + \lambda \int_R \pi_t(u) \ln \pi_t(u) du,
\]

where \( \dot{V}_t^\pi = \frac{V^\pi(X_{t+\Delta t}, t+\Delta t) - V^\pi(X_t, t)}{\Delta t} \) is the total derivative and \( \Delta t \) is the discretization step for the learning algorithm.

The objective of the policy evaluation procedure is to minimize the Bellman’s error \( \delta_t \). In general, this can be carried out as follows. Denote by \( V^\theta \) and \( \pi^\phi \) respectively the parametrized value function and policy (upon using regressions or neural networks, or making use of some of the structures of the problem; see below), with \( \theta, \phi \) being the vector of weights to be learned. We then minimize

\[
C(\theta, \phi) = \frac{1}{2} E \left[ \int_0^T |\delta_t|^2 dt \right] = \frac{1}{2} E \left[ \int_0^T |\dot{V}_t^\theta + \lambda \int_R \pi^\phi_t(u) \ln \pi^\phi_t(u) du|^2 dt \right],
\]

where \( \pi^\phi = \{\pi^\phi_t, t \in [0, T]\} \) is generated from \( \pi^\phi \) with respect to a given initial state \( X_0 = x \) at time 0. To approximate \( C(\theta, \phi) \) in an implementable algorithm, we first discretize \([0, T]\) into small equal-length intervals \([t_i, t_{i+1}]\), \( i = 0, 1, \ldots, l \) where \( t_0 = 0 \) and \( t_{l+1} = T \). Then we collect a set of samples
\( D = \{(t_i, x_i), i = 0, 1, \cdots, l + 1\} \) in the following way. The initial sample \((t_0, x_0) = (0, x)\). Now, at each \(t_i, i = 1, \cdots, l\), we sample \(\pi_i^\phi\) to obtain an allocation \(u_i \in \mathbb{R}\) in the risky asset, and then calculate the wealth \(x_{i+1}\) at the next time instant \(t_{i+1}\) according to the discretized version of the classical controlled wealth dynamics (2):

\[
x_{i+1} = x_i + \sigma u_i (\rho \Delta t + W_{t_{i+1}} - W_{t_i})
\]

We can now approximate \(C(\theta, \phi)\) by

\[
C(\theta, \phi) = \frac{1}{2} \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) + \lambda \int_{\mathbb{R}} \pi_i^\phi(u) \ln \pi_i^\phi(u) du \right)^2 \Delta t. \quad (43)
\]

Instead of following the common practice to represent \(V^\theta, \pi^\phi\) using (deep) neural networks for continuous RL problems, in this paper we will take advantage of the more explicit parametric expressions stipulated in Theorem 1 and Theorem 6. This will lead to faster learning and convergence, which will be demonstrated in the experiments below. More precisely, by virtue of Theorem 6, we will focus on Gaussian policies with variance taking the form

\[
\sigma^2 = \lambda \pi e^{1-2\phi_1} \quad \text{and} \quad \theta_3 = 2\phi_2.
\]

The improved policy, in turn, becomes, according to (32), that

\[
\pi(u; t, x, w) = N \left( u \mid - \frac{\theta}{\sigma^2} (x - w), \frac{\lambda}{2\sigma^2} e^{\theta_3(T-t)} \right) \]

\[
= N \left( u \mid - \sqrt{\frac{2\phi_2}{\lambda \pi}} e^{\frac{2\phi_1-1}{2}} (x - w), \frac{1}{2\pi} e^{2\phi_2(T-t)+2\phi_1-1} \right), \quad (46)
\]

On the other hand, as suggested by the theoretical optimal value function (23) in Theorem 1, we consider the parameterized \(V^\theta\), where \(\theta = (\theta_0, \theta_1, \theta_2, \theta_3)'\), by

\[
V^\theta(t, x) = (x - w)^2 e^{-\theta_3(T-t)} + \theta_2 t^2 + \theta_1 t + \theta_0, \quad (t, x) \in [0, T] \times \mathbb{R}. \quad (44)
\]
where we have assumed that the true (unknown) Sharpe ratio \( \rho > 0 \).

Rewriting the objective (43) using \( H(\pi_t^\phi) = \phi_1 + \phi_2(T - t) \), we obtain

\[
C(\theta, \phi) = \frac{1}{2} \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)) \right)^2 \Delta t.
\]

Note that \( \dot{V}^\theta(t_i, x_i) = \frac{V^\theta(t_{i+1}, x_{i+1}) - V^\theta(t_i, x_i)}{\Delta t} \), with \( \theta_3 = 2\phi_2 \) in the parametrization of \( V^\theta(t_i, x_i) \). It is now straightforward to devise the updating rules for \( \theta_1, \theta_2 \) and \( \phi_1, \phi_2 \) using stochastic gradient descent algorithms (see, for example, Chapter 8 of Goodfellow et al. (2016)). Precisely, we compute

\[
\frac{\partial C}{\partial \phi_2} = \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)) \right) \Delta t \times \left( -\frac{2(x_{i+1} - w)^2e^{-2\phi_2(T - t_{i+1})} - 2(x_i - w)^2e^{-2\phi_2(T - t_i)} - \lambda(T - t_i)}{\Delta t} \right);
\]

\[
\frac{\partial C}{\partial \phi_1} = -\lambda \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)) \right) \Delta t; \quad (47)
\]

\[
\frac{\partial C}{\partial \theta_2} = \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)) \right) \left( t_{i+1}^2 - t_i^2 \right); \quad (49)
\]

\[
\frac{\partial C}{\partial \theta_1} = \sum_{(t_i, x_i) \in D} \left( \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)) \right) \Delta t. \quad (50)
\]

Moreover, the parameter \( \theta_3 \) is updated with \( \theta_3 = 2\phi_2 \), and \( \theta_0 \) is updated based on the terminal condition \( V^\theta(T, x; w) = (x - w)^2 - (w - z)^2 \), which yields

\[
\theta_0 = -\theta_2T^2 - \theta_1T - (w - z)^2. \quad (51)
\]

Finally, we provide a scheme for learning the underlying Lagrange multiplier \( w \). Indeed, the constraint \( \mathbb{E}[X_T] = z \) itself suggests the standard stochastic approximation update

\[
w_{n+1} = w_n - \alpha_n(X_T - z), \quad (52)
\]

with \( \alpha_n > 0, n \geq 1 \), being the learning rate. In implementation, we can replace \( X_T \) in (52) by a sample average \( \frac{1}{N} \sum_i x_T^i \) to have a more stable learning process (see, for example, Section 1.1 of Kushner and Yin (2003)), where \( N \geq 1 \) is the sample size and \( x_T^i \) is the terminal wealth generated by the \( i \)-th
iteration of the learning process. It is interesting to notice that the learning scheme of $w$, (52), is statistically self-correcting. For example, if the (sample average) terminal wealth is above the target $z$, the updating rule (52) will decrease $w$, which in turn decreases the mean of the exploratory Gaussian policy $\pi$ in view of (46). This implies that there will be less risky allocation in the next step of actions for learning (exploration) and optimizing (exploitation), leading to, on average, a decreased terminal wealth.

We now summarize the pseudocode for the ENT-MV algorithm in Algorithm 1.

\begin{algorithm}
\caption{ENT-MV: Entropy-Regularized Mean-Variance Portfolio Optimization Reinforcement Learning}
\begin{algorithmic}
\Input{Market Simulator $Market$, learning rates $\alpha, \eta_\theta, \eta_\phi$, investment horizon $T$, discretization $\Delta t$, exploration rate $\lambda$, number of iterations $M$, sample average size $N$.}
\State Initialize $\theta$, $\phi$ and $w$ for $k = 1$ to $M$
\For{$k = 1$ to $M$}
\For{$i = 1$ to $\lfloor \frac{T}{\Delta t} \rfloor$}
\Sample{$(t_i^k, x_i^k)$} from $Market$
\Obtain collected samples $D = \{(t_i^k, x_i^k), 1 \leq i \leq \lfloor \frac{T}{\Delta t} \rfloor\}$
\Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta C(\theta, \phi)$ using (50) and (49)
\Update $\theta_0$ using (51)
\Update $\phi \leftarrow \phi - \eta_\phi \nabla_\phi C(\theta, \phi)$ using (48) and (47)
\EndFor
\Update $\pi^\phi \leftarrow \mathcal{N}\left( u \mid -\sqrt{\frac{2\phi_2}{\pi^2}} e^{\frac{2\phi_1-1}{2}(x - w)} \cdot \frac{1}{2\pi} e^{\frac{2\phi_2(T-t)+2\phi_1-1}{2}} \right)$
\If{$k \mod N == 0$}
\Update $w \leftarrow w - \alpha \left( \frac{1}{N} \sum_{j=k-N+1}^{k} x_j^\phi \lfloor \frac{T}{\Delta t} \rfloor - z \right)$
\EndIf
\EndFor
\end{algorithmic}
\end{algorithm}

5 Simulation Results

In this section, we compare the performance of our RL algorithm ENT-MV with two other methods that could be used to solve the classical MV problem (3). The first one is the traditional maximum likelihood estimation (MLE)

\footnote{This discussion is based on the assumption that the market has a positive Sharpe ratio (recall (46)). The case of a negative Sharpe ratio can be dealt with similarly.}
that relies on the real-time estimation of the drift $\mu$ and the volatility $\sigma$ in the geometric Brownian motion model for the price process (1). Once the estimators of $\mu$ and $\sigma$ are available using the most recent price time series, the allocation can be computed using the optimal allocation (28) for the classical MV problem. Another alternative is based on the deep deterministic policy gradient (DDPG) method developed by Lillicrap et al. (2016), a method that has been taken as the baseline approach for comparisons of different RL algorithms that solve continuous control problems.

We conduct comparisons in various simulation settings, including the stationary setting where the price of the risky asset evolves according to the geometric Brownian motion (1) and the non-stationary setting that involves stochastic factor modeling for the drift and volatility parameters. We also consider a time-decreasing exploration case which can be easily accommodated by the ENT-MV through an annealing temperature parameter $\lambda$. In almost all the experiments, our ENT-MV algorithm outperforms the other two methods by large margins.

In the following, we briefly describe the two other methods.

**Maximum likelihood estimation (MLE)**
The MLE is a popular method for estimating the parameters $\mu$ and $\sigma$ in the geometric Brownian motion model (1). We refer the interested readers to Section 9.3.2 of Campbell et al. (1997) for a detailed description of this method. At each decision making time $t_i$, the MLE estimators for $\mu$ and $\sigma$ are calculated based on the most recent 100 data points of the price. One can then substitute the MLE estimators into the optimal allocation (28) and the expression of the Lagrange multiplier $w = \frac{e^{\rho T} - \mu}{\sigma^2 \rho^2} - 1$ to compute the allocation $u_i \in \mathbb{R}$ in the risky asset. Such a two-phase procedure is commonly used in adaptive control, where the first phase is known as the adaptation phase and the second one is the optimization phase (see, for example, Chen and Guo (2012), Kumar and Varaiya (2015)). The real-time estimation procedure also allows the MLE to be applicable in non-stationary markets with time-varying $\mu$ and $\sigma$.

**Deep deterministic policy gradient (DDPG)**
The DDPG method has attracted significant attention since it was introduced in Lillicrap et al. (2016). It has been serving as a state-of-the-art baseline approach for continuous control (action) RL problems, albeit in discrete time. The DDPG learns a deterministic target policy using deep neural networks for both the critic and the actor, with exogenous noise being added to encourage exploration (e.g. using OU processes; see Lillicrap et al. (2016)).
(2016) for details). To adapt the DDPG to the classical MV setting (without entropy regularization), we make the following adjustments. Since the target policy we aim to learn is a deterministic function of \((x - w)\) (cf. (28)), we will feed the samples \(x_i - w\), rather than only \(x_i\), to the actor network in order to output the allocation \(u_i \in \mathbb{R}\). Here, \(w\) is the learned Lagrange multiplier at the decision making time \(t_i\), obtained from the same self-correcting scheme (52). This modification also enables us to connect current allocation with the previously obtained sample average of terminal wealth in a feedback loop, through the Lagrange multiplier \(w\). Another change from the original DDPG is that we include prioritized experience replay (Schaul et al. (2016)), rather than sampling experience uniformly from the replay buffer. We select the terminal experience with higher probability to train the critic and actor networks, to account for the fact that the MV problem has no running cost, but only a terminal cost given by \((x_T - w)^2 - (w - z)\) (cf. (4)). Such a modification significantly improves learning speed and performance.

5.1 The stationary market case

We first perform numerical simulations in the stationary market environment, where the price process is simulated according to the geometric Brownian motion (1) with constant \(\mu\) and \(\sigma\). We take \(T = 1\) and \(\Delta t = \frac{1}{252}\), indicating that the MV problem is considered over a one-year period, with daily rebalancing. Reasonable values of the annualized return and volatility will be taken from the sets \(\mu \in \{-50\%, -30\%, -10\%, 0\%, 10\%, 30\%, 50\%\}\) and \(\sigma \in \{10\%, 20\%, 30\%, 40\%\}\), respectively. These values are usually considered for a “typical” stock for simulation purpose (see, for example, Hutchinson et al. (1994)). The annualized interest rate is taken to be \(r = 2\%\). We consider the MV problem with a 40\% annualized target return on the terminal wealth starting from a normalized initial wealth \(x = 1\) and, hence, \(z = 1.4\) in (3). These model parameters will be fixed for all the simulations considered in this paper.

For the ENT-MV algorithm, we take the total training episodes \(M = 20000\) and the sample size \(N = 10\) for learning the Lagrange multiplier \(w\). The temperature parameter \(\lambda = 2\). Across all the simulations in this section, the learning rates are fixed as \(\alpha = 0.05\) and \(\eta_{\theta} = \eta_{\phi} = 0.0005\).

We choose the same \(M\) and \(N\) for the DDPG algorithm for fair comparison. The critic network has 3 hidden layers and with 10, 8, 8 hidden units for each layer, and the actor network has 2 hidden layers with 10, 8 hidden units. The learning rates for critic and actor are 0.0001, fixed across all simulations in this section as the case for ENT-MV. The replay buffer has
size 80, and the minibatch size is 20 for the stochastic gradient decent. The target network has the soft update parameter $\tau = 0.001$. Finally, we adopt the OU process for adding exploration noise; see Lillicrap et al. (2016) for more details. All the simulations using the DDPG algorithm were trained in Tensorflow.

We summarize in Table 1 the simulation results for the three methods, ENT-MV, MLE and DDPG, under market scenarios corresponding to different combinations of $\mu$’s and $\sigma$’s. For each method under each market scenario, we present the annualized sample mean $\overline{M}$ and sample variance $\overline{V}$ of the last 2000 terminal wealth, and the corresponding annualized Sharpe ratio $SR = \frac{\overline{M} - 1}{\sqrt{\overline{V}}}$.

A few observations can be made based on the simulation results in Table 1. First of all, the ENT-MV algorithm outperforms the other two methods by a large margin, in terms of several statistics of the investment outcome including sample mean, sample variance and Sharpe ratio. In fact, based on the comparison of Sharpe ratios, the ENT-MV algorithm outperforms the MLE method over all the 28 experiments, and outperforms the DDPG algorithm in 23 out of the total 28 experiments. Notice that the DDPG method yields rather unstable performance across different market scenarios, with some of the sample average terminal wealth below 0 indicating bankruptcy. The ENT-MV algorithm, on the other hand, achieves positive annualized return over all the experiments. The superiority of the ENT-MV algorithm over the deep RL method DDPG is even paramount if we take into account the training time (all experiments were performed on an MacBook Air laptop). The DDPG method involves extensive training of two deep neural networks, making it less appealing for high-frequency portfolio rebalancing and trading in practice.

Another advantage of ENT-MV over DDPG is the ease of hyperparameter tuning. Recall that the learning rates are fixed respectively for ENT-MV and DDPG across all the experiments for different $\mu$ and $\sigma$. The performance of the ENT-MV algorithm is less affected by the fixed learning rates, while the DDPG algorithm is not. Indeed, DDPG has been noted for its notorious brittleness and hyperparameter sensitivity in practice (see, for example, Duan et al. (2016), Henderson et al. (2018)), issues also shared by other deep RL methods for continuous control problems. Our ENT-MV algorithm does not suffer from such issues, as a consequence of avoiding deep neural networks in its training and decision making processes.

The MLE method, although free of any learning rates tuning, directly estimates the underlying parameters $\mu$ and $\sigma$. Across all the simulations
Table 1: Comparison of the annualized sample mean ($\overline{M}$), sample variance ($\overline{V}$), Sharpe ratio ($SR$) and average training time (per experiment) for ENT-MV, MLE and DDPG.

| Market scenarios | ENT-MV  | MLE     | DDPG    |
|------------------|---------|---------|---------|
| $\mu = -50\%, \sigma = 10\%$ | 1.396; 0.006; 5.107 | 1.556; 0.017; 4.284 | 1.297; 0.107; 0.908 |
| $\mu = -30\%, \sigma = 10\%$ | 1.390; 0.016; 3.039 | 1.215; 0.014; 1.833 | 1.401; 0.003; 7.076 |
| $\mu = -10\%, \sigma = 10\%$ | 1.330; 0.074; 1.218 | 1.056; 1.365; 0.482 | 0.901; 0.014; -0.833 |
| $\mu = 0\%, \sigma = 10\%$ | 1.204; 1.280; 0.180 | 0.926; 38.40; -0.012 | 1.029; 0.038; 0.147 |
| $\mu = 10\%, \sigma = 10\%$ | 1.318; 0.171; 0.769 | 1.009; 0.453; 0.014 | 0.951; 0.008; -0.541 |
| $\mu = 30\%, \sigma = 10\%$ | 1.385; 0.019; 2.785 | 1.179; 0.059; 0.737 | 0.224; 0.104; -2.405 |
| $\mu = 50\%, \sigma = 10\%$ | 1.394; 0.007; 4.772 | 1.459; 0.013; 3.983 | 1.478; 0.667; 0.717 |
| $\mu = -50\%, \sigma = 20\%$ | 1.387; 0.007; 4.772 | 1.459; 0.013; 3.983 | 1.478; 0.667; 0.717 |
| $\mu = -30\%, \sigma = 20\%$ | 1.359; 0.051; 1.598 | 1.205; 1.319; 0.178 | 1.973; 0.404; 1.531 |
| $\mu = -10\%, \sigma = 20\%$ | 1.309; 0.245; 0.625 | 1.036; 2.183; 0.024 | 1.349; 0.368; 0.575 |
| $\mu = 0\%, \sigma = 20\%$ | 1.105; 0.727; 0.123 | 0.921; 6.887; -0.301 | 0.988; 0.139; -0.033 |
| $\mu = 10\%, \sigma = 20\%$ | 1.221; 0.314; 0.395 | 1.045; 17.87; -0.023 | 1.243; 0.534; 0.408 |
| $\mu = 30\%, \sigma = 20\%$ | 1.345; 0.062; 1.387 | 1.155; 1.743; 0.117 | 1.360; 0.050; 1.613 |
| $\mu = 50\%, \sigma = 20\%$ | 1.385; 0.027; 2.350 | 1.237; 1.293; 0.208 | 1.385; 0.004; 6.496 |
| $\mu = -50\%, \sigma = 30\%$ | 1.353; 0.044; 1.682 | 1.333; 9.465; 0.108 | 0.272; 2.762; -0.438 |
| $\mu = -30\%, \sigma = 30\%$ | 1.323; 0.106; 0.992 | 1.092; 5.657; 0.039 | 0.034; 0.924; -1.005 |
| $\mu = -10\%, \sigma = 30\%$ | 1.317; 0.696; 0.380 | 1.045; 17.87; 0.011 | 1.371; 0.792; 0.417 |
| $\mu = 0\%, \sigma = 30\%$ | 1.079; 0.727; 0.092 | 0.955; 28.84; -0.008 | 1.070; 0.752; 0.081 |
| $\mu = 10\%, \sigma = 30\%$ | 1.282; 0.885; 0.300 | 0.885; 24.06; -0.023 | 1.243; 0.825; 0.268 |
| $\mu = 30\%, \sigma = 30\%$ | 1.334; 0.131; 0.921 | 0.886; 24.41; -0.023 | 1.210; 0.921; 0.218 |
| $\mu = 50\%, \sigma = 30\%$ | 1.350; 0.049; 1.583 | 1.238; 7.505; 0.087 | 0.610; 0.143; -1.030 |
| $\mu = -50\%, \sigma = 40\%$ | 1.342; 0.061; 1.385 | 1.284; 11.14; 0.085 | 1.328; 0.501; 0.463 |
| $\mu = -30\%, \sigma = 40\%$ | 1.320; 0.146; 0.839 | 1.145; 3.315; 0.080 | 1.212; 0.160; 0.531 |
| $\mu = -10\%, \sigma = 40\%$ | 1.241; 0.707; 0.287 | 0.979; 7.960; -0.007 | 1.335; 1.413; 0.282 |
| $\mu = 0\%, \sigma = 40\%$ | 1.057; 0.671; 0.070 | 0.950; 31.60; -0.009 | 1.064; 1.467; 0.053 |
| $\mu = 10\%, \sigma = 40\%$ | 1.155; 0.591; 0.202 | 1.053; 9.090; 0.017 | 1.242; 1.499; 0.198 |
| $\mu = 30\%, \sigma = 40\%$ | 1.320; 0.198; 0.716 | 1.083; 17.46; 0.020 | 0.179; 1.533; -0.663 |
| $\mu = 50\%, \sigma = 40\%$ | 1.329; 0.078; 1.174 | 0.963; 43.17; -0.006 | -0.390; 1.577; -1.107 |

**Training time**

| ENT-MV | MLE | DDPG |
|--------|-----|------|
| $<10s$ | $<10s$ | $\approx 3hrs$ |

---

14 If the hyperparameters are allowed to be tuned on a scenario-by-scenario basis (i.e., not for comparison purpose), the performance of ENT-MV can be further improved with the Sharpe ratio close to its theoretical maximum, whereas improvement is more difficult to achieve for DDPG by hyperparameter tuning.
presented in Table 1, the estimated value of $\sigma$ is relatively close to its truth, while the drift parameter $\mu$ cannot be estimated accurately. This is consistent with the mean–blur problem, which in turn leads to higher variance of the terminal wealth (see Table 1) when one applies the estimated $\mu$ and $\sigma$ to select the allocation (28) in the risky asset.

Before concluding this section, we also plot the learning curves for the three methods in Figure 1 and Figure 2. These learning curves display the change of the sample mean and sample variance of every (non-overlapping) 50 terminal wealth, as the learning continues for each method. Notice the log scale for the variance in Figure 2. From these plots, we can see that the ENT-MV algorithm converges relatively faster than the other two methods, achieving relatively good performance even in the front phase of the learning process. This is also consistent with the convergence result in Theorem 6 and the remarks following it in section 4.1.

![Learning curves of sample means of terminal wealth](image)

**Figure 1:** Learning curves of sample means of terminal wealth (over every 50 iterations) for ENT-MV, MLE and DDPG ($\mu = -30\%, \sigma = 10\%$).
5.2 The non-stationary market case

One of the major differences between RL applications in quantitative finance and other areas is that the underlying unknown investment environment is typically time-varying. In this section, we consider the performance of the three methods in the non-stationary scenario where the price process is modeled by the stochastic factor model. To have a well defined learning problem, the stochastic factor should change at a much slower time-scale, compared to the scale of the learning process. Specifically, we adopt the slow factor model in the multi-scale stochastic volatility model framework (see, for example, Fouque et al. (2003)). The price process follows

\[ dS_t = S_t(\mu_t \, dt + \sigma \, dW_t), \quad 0 < t \leq T, \quad S_0 = s > 0, \]

with \( \mu_t, \sigma_t, t \in [0, T] \) being the drift and volatility processes restricted to each episode \([0, T]\). The controlled wealth dynamics is therefore

\[ dx_t^u = \sigma_t u_t(\rho_t \, dt + dW_t), \quad 0 < t \leq T, \quad x_0^u = x \in \mathbb{R}. \quad (53) \]

For simulations in this section, we take

\[ d\rho_t = \delta dt, \quad \text{and} \quad d\sigma_t = \sigma_t(\delta dt + \sqrt{\delta} \, dW_t^1), \quad 0 < t \leq MT, \quad (54) \]
with $\rho_0 \in \mathbb{R}$ and $\sigma_0 > 0$ being the initial values, respectively, $\delta > 0$ being a small parameter, and the two Brownian motions $d\langle W, W^1 \rangle_t = \gamma dt$, with $|\gamma| < 1$. Notice that the terminal horizon for the stochastic factor model is $MT$ (recall $M = 20000$ and $T = 1$ as in section 5.1), which indicates that the stochastic factors $\rho_t$ and $\sigma_t$ change across all the $M$ episodes. To make sure that their values stay in reasonable ranges after running for all $M$ episodes, we take $\delta$ to be small with value $\delta = 0.0001$, and $\rho_0 = -3.2$, $\sigma_0 = 10\%$, corresponding to the case $\mu_0 = -30\%$ initially. We plot the learning curves in Figure 3 and Figure 4. The ENT-MV algorithm displays remarkable stability for learning performance, compared to the other two methods even in the non-stationary market scenario.

Figure 3: Learning curves of sample means of terminal wealth (over every 50 iterations) for ENT-MV, MLE and DDPG for non-stationary market scenario ($\mu_0 = -30\%, \sigma_0 = 10\%, \delta = 0.0001, \gamma = 0$).
Figure 4: Learning curves of sample variances of terminal wealth (over every 50 iterations) for ENT-MV, MLE and DDPG for non-stationary market scenario ($\mu_0 = -30\%, \sigma_0 = 10\%, \delta = 0.0001, \gamma = 0$).

5.3 The decaying exploration case

A decaying exploration scheme is often desirable for RL problems as more iterations have been run. It suggests that exploitation gradually takes more weight over exploration towards the end of the learning process. Within the current entropy-regularized relaxed stochastic control framework, we have been able to show the convergence from the solution of the exploratory MV to that of the classical MV in Theorem 3 as the trade-off parameter $\lambda \to 0$. Putting together Theorem 3 and Theorem 6 we can reasonably expect a further improvement of the ENT-MV algorithm when it adopts a decaying $\lambda$ scheme, rather than a constant $\lambda$ case as in the previous two sections. Our numerical simulation result in Figure 5 demonstrates slightly improved performance of the ENT-MV algorithm with a specifically chosen $\lambda$ process that decreases across all the $M$ episodes, given by

$$\lambda_k = \lambda_0 \left(1 - \exp\left(\frac{200(k - M)}{M}\right)\right), \quad \text{for} \quad k = 0, 1, 2, \ldots, M.$$  \hfill (55)

We plot the histogram of the last 2000 wealth generated by the ENT-MV algorithm with the decaying $\lambda$ scheme starting from $\lambda_0 = 2$ and the original
ENT-MV algorithm with constant $\lambda = 2$, respectively. The Sharpe ratio increases from 3.039 to 3.243. The result in Figure 5 corresponds to the stationary market case; the case for non-stationary market demonstrates similar improvement by adopting the decaying exploration scheme.

![Histogram of last 2000 terminal wealth for ENT-MV with constant $\lambda$ and decaying $\lambda$ ($\mu = -30\%, \sigma = 10\%$).](image)

Figure 5: Histogram of last 2000 terminal wealth for ENT-MV with constant $\lambda$ and decaying $\lambda$ ($\mu = -30\%, \sigma = 10\%$).

6 Conclusions

We consider solving the continuous-time MV problem under the RL framework, by adopting the recently developed framework for entropy-regularized relaxed stochastic control problems in Wang et al. (2019). Our method recast the MV portfolio optimization question as a learning and exploration problem, rather than merely an estimation question of the unknown model parameters that is commonly accepted to be challenging in investment practice. The exploration phase is naturally captured by both the objective function and the exploratory state dynamics of the new optimization problem. We derive that the optimal feedback control distribution that balance exploration and exploitation in the MV setting is a Gaussian distribution with time-decaying variance. Similar as the case for general LQ problems in Wang et al. (2019), we establish the close connections between
the exploratory MV and the classical MV problems, including the solvability equivalence and the convergence as exploration decays.

The reformulation of the classical MV problem also allows us to design a competitive RL algorithm, thanks in no small measure to a provable policy improvement theorem and the explicit functional structures for both the value function and the policy. The policy improvement theorem yields an updating scheme for the policy in the continuous-time stochastic control setting with both entropy regularization and control relaxation. On the other hand, the explicit structures of the theoretical optimal solution to the exploratory MV problem give rise to simple yet efficient function approximators without using black-box neural network approximations. The advantage of our method is evidently shown in various numerical simulations under both the stationary and non-stationary market environment cases, where our RL algorithm generally outperforms other state-of-the-art approaches by large margins when solving the MV problem in the continuous control setting.

Interesting future research directions include empirical study of our RL algorithm using real data, in the high-dimensional control setting where allocations across multiple risky assets are determined at each decision making time. For that, our exploratory MV formulation still works and a multivariate Gaussian distribution would be learned using an algorithm similar to Algorithm 1. It is interesting to compare the performance of our algorithm with other existing methods, including the Fama-French model (Fama and French (1993)) and the distributionally robust method for MV problem (Blanchet et al. (2018)). Another open question is to design an “optimal” decaying scheme for the temperature parameter $\lambda$, an essential quantity that bridges the exploratory MV problem and the classical MV problem. These questions are left for further investigation.

References

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

Tomasz R Bielecki, Haoning Jin, Stanley R Pliska, and Xun Yu Zhou. Continuous-time mean-variance portfolio selection with bankruptcy prohibition. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 15(2):213–244, 2005.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-
variance portfolio selection with wasserstein distances. *arXiv preprint arXiv:1802.04885*, 2018.

John Y Campbell, Andrew W Lo, and A Craig MacKinlay. *The econometrics of financial markets* princeton University press, 1997.

Han-Fu Chen and Lei Guo. *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012.

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

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

Darrell Duffie and Henry R Richardson. Mean-variance hedging in continuous time. *The Annals of Applied Probability*, 1(1):1–15, 1991.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

Jean-Pierre Fouque, George Papanicolaou, Ronnie Sircar, and Knut Solna. Multiscale stochastic volatility asymptotics. *Multiscale Modeling & Simulation*, 2(1):22–42, 2003.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. [http://www.deeplearningbook.org](http://www.deeplearningbook.org).

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1352–1361, 2017.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Dieter Hendricks and Diane Wilcox. A reinforcement learning extension to the almgren-chriss framework for optimal trade execution. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 457–464. IEEE, 2014.
James M Hutchinson, Andrew W Lo, and Tomaso Poggio. A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, 49(3):851–889, 1994.

Saul D Jacka and Aleksandar Mijatović. On the policy improvement algorithm in continuous time. *Stochastics*, 89(1):348–359, 2017.

Panqnamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*, volume 75. SIAM, 2015.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Duan Li and Wan-Lung Ng. Optimal dynamic portfolio selection: Multi-period mean-variance formulation. *Mathematical finance*, 10(3):387–406, 2000.

Xun Li, Xun Yu Zhou, and Andrew EB Lim. Dynamic mean-variance portfolio selection with no-shorting constraints. *SIAM Journal on Control and Optimization*, 40(5):1540–1555, 2002.

Timothy Lillicrap, Jonathan Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Andrew EB Lim and Xun Yu Zhou. Mean-variance portfolio selection with random parameters in a complete market. *Mathematics of Operations Research*, 27(1):101–120, 2002.

David G Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

Shie Mannor and John N Tsitsiklis. Algorithmic aspects of mean–variance optimization in markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.

Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
John Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE transactions on neural Networks*, 12(4):875–889, 2001.

John Moody, Lihzhong Wu, Yuansong Liao, and Matthew Saffell. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5-6):441–470, 1998.

Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299, 2000.

Rémi Munos and Paul Bourgine. Reinforcement learning for continuous stochastic control problems. In *Advances in neural information processing systems*, pages 1029–1035, 1998.

Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680. ACM, 2006.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In *Advances in neural information processing systems*, pages 252–260, 2013.

LA Prashanth and Mohammad Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward mdps. *Machine Learning*, 105(3):367–417, 2016.

Makoto Sato and Shigenobu Kobayashi. Variance-penalized reinforcement learning for risk-averse asset allocation. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 244–249. Springer, 2000.

Makoto Sato, Hajime Kimura, and Shibenoobu Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, Puerto Rico, 2016.

Matthew J Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.
Robert Henry Strotz. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180, 1955.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Aviv Tamar and Shie Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Temporal difference methods for the variance of the reward to go. In *ICML (3)*, pages 495–503, 2013.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Exploration versus exploitation in reinforcement learning: A stochastic control approach. *arXiv preprint: arXiv:1812.01552v3*, 2019.

Xun Yu Zhou and Duan Li. Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, 42(1):19–33, 2000.

Xun Yu Zhou and George Yin. Markowitz’s mean-variance portfolio selection with regime switching: A continuous-time model. *SIAM Journal on Control and Optimization*, 42(4):1466–1482, 2003.