Dynamic mean–variance portfolio selection with no-shorting constraints and unknown investment opportunity sets

Xun Li* Yutian Wang† Xun Yu Zhou‡

July 18, 2026

Abstract

We study continuous-time mean–variance portfolio selection with no-shorting constraints and unknown investment opportunity sets from a reinforcement learning (RL) perspective. The problem is a *constrained* stochastic linear–quadratic control problem for which the entropy-regularized exploratory formulation of Wang et al. (2020) leads to difficulty in theoretical analysis, because enforcing the constraint on the support of randomized policies nullifies the tractable Gaussian exploration. To tackle this challenge, we introduce an auxiliary exploratory problem *without* entropy in which exploratory policies are still Gaussian whose samples may violate the no-shorting requirement but their means satisfy it. We then prove that, for a suitable choice of exploration variance, the mean of the optimal Gaussian policy of the auxiliary problem coincides with the optimal policy of the original problem. Motivated by this theoretical result, we develop a model-free RL algorithm that learns the optimal policy of the auxiliary (and hence the original) problem directly from trajectory data without estimating the investment opportunity set. A numerical example demonstrates the performance of the proposed algorithm.

Keywords: Mean–variance portfolio selection; no-shorting; reinforcement learning; Gaussian exploration; policy evaluation; policy gradient.

1 Introduction

Mean–variance (MV) portfolio selection introduced by Markowitz (1952) is a cornerstone of the modern quantitative finance theory and practice. The original model of Markowitz (1952) is static, and there has been a sizable literature developed over the last 26 years on the continuous-time extension of the MV theory, starting from Zhou and Li (2000). The main idea of Zhou and Li (2000) is to turn continuous-time MV into an *indefinite* stochastic linear–quadratic (LQ) control problem, whose theory was developed in the 1990s (Chen et al. 1998), leading to explicit characterization of the efficient frontier. One of the focuses of the subsequent study of continuous-time MV is to incorporate various portfolio constraints such as those of no-shorting and no-bankruptcy. Those constraints translate into control and state constraints, which introduce great technical difficulties because the general LQ control theory does not cope with any constraints. New and ad hoc approaches have been developed exploiting the special features of the MV problems, including that the state (wealth) is scalar-valued. Li et al.

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. E-mail: li.xun@polyu.edu.hk

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. E-mail: yutian.wang@connect.polyu.hk

‡Department of Industrial Engineering and Operations Research & Data Science Institute, Columbia University, New York, USA, NY 10027. E-mail: xz2574@columbia.edu

(2002) solve continuous-time MV portfolio selection when shorting is disallowed, by introducing two Riccati equations and a viscosity solution analysis of the associated HJB equation. Hu and Zhou (2005) further develop a general theory on cone-constrained stochastic LQ control with random coefficients, obtain explicit optimal controls via two extended stochastic Riccati equations and apply the results to MV under no-shorting constraints. Li and Xu (2016) investigate a continuous-time Markowitz model with bankruptcy prohibition and convex cone portfolio constraints, and derive semi-analytical solutions via a martingale approach without relying on viscosity solutions.

The study on continuous-time MV prior to the year 2020 had been entirely model-based, namely, under the assumption that the investment opportunity set (consisting of the asset dynamics coefficients) is given and accurate. However, it is well understood that it is hard, if not impossible, to accurately estimate the moments of asset returns, especially the first moment, based on price data which are typically very limited. On the other hand, portfolios and their objective values derived from analytical/numerical solutions of the MV problems are known to be extremely sensitive to those estimation errors. These render the derived model-based portfolio policies often irrelevant or even misleading.

This is where reinforcement learning (RL) comes to rescue. RL studies how an agent learns to make sequential decisions from interaction with an *unknown* environment. Primarily premised upon dynamic programming through value-based and temporal-difference methods, the RL theory and algorithms have been developed mainly for problems in discrete time with discrete state and action (control) spaces; see, for example, the classical textbook Sutton and Barto (2018). Wang et al. (2020) are the first to formulate the continuous-time RL with possibly continuous state/action spaces, using relaxed controls and entropy regularization to model exploration when the model parameters are unknown. In particular, they prove theoretically that in the LQ setting the optimal exploratory feedback policy must be Gaussian. This provides the interpretability of the Gaussian exploration commonly employed in practice, which is easy to sample from, differentiate and optimize. Wang and Zhou (2020) immediately apply this general theory to the continuous-time MV portfolio selection problem with an unknown investment opportunity set to prove that the optimal exploratory policy is Gaussian with a time-decaying variance, and establish a policy improvement theorem leading to an implementable RL algorithm. Subsequently, Jia and Zhou (2022a,b, 2023) develop a systematic theory on policy evaluation and temporal-difference learning, policy gradient and actor–critic methods, and continuous-time q-learning, respectively. Tang et al. (2022) establish the well-posedness, regularity, and vanishing-exploration convergence for the exploratory HJB equations. More recently, Huang et al. (2025) investigate the convergence of policy iteration for general entropy-regularized stochastic control problems, and Sethi et al. (2025) study related algorithmic and asymptotic properties of continuous-time policy mirror descent with entropy annealing. Guo et al. (2026) propose a regularized policy gradient method for infinite-horizon LQ control and establish linear and partly superlinear convergence. For the continuous-time MV problem, Huang et al. (2024) develop a model-free, data-driven RL algorithm based on the policy gradient theory of Jia and Zhou (2022b), prove not only its convergence but also a *sublinear* regret bound in terms of the Sharpe ratio, and carry out an extensive empirical study to demonstrate its superior performance.

The relaxed or exploratory controls are in general probability distributions on action/control spaces. If action spaces are unconstrained, then the resulting exploratory problems are more likely to have better analytical structures leading to more efficient learning algorithms. For example, the optimality of Gaussian exploration proved in Wang et al. (2020), Wang and Zhou (2020) and Huang et al. (2024) stems not only from the underlying LQ framework but also from the unconstrained action spaces. When there are constraints the exploratory controls will be supported on the constrained action space. This likely leads to *truncated* Gaussian-type exploratory policies for the current problem of MV portfolio selection without shorting. Indeed, in a different setting of the Merton problem, Chau et al. (2026) show that the optimal exploratory policy is truncated Gaussian under interval constraints. Such a formulation ensures that *every* control sampled from the exploratory policy is feasible, which however

comes with a major drawback: The HJB equation and the corresponding learning problem become substantially harder to analyze, and the explicit structure available under Gaussian exploration is generally lost.

To overcome this difficulty, we approach the constrained problem by imposing the constraint on the policy *mean* rather than on the support. More precisely, we keep any exploratory policy Gaussian, but require its mean to lie in the positive half-space. Hence the randomized policy may place some mass outside $\mathbb{R}^m_+$, while its "averaged" policy is componentwise nonnegative and therefore admissible for the original problem. We thus formulate an auxiliary exploratory problem where the class of admissible policies are those Gaussian distributions with nonnegative means and a suitable choice of variance. To justify the relevance of this auxiliary problem vis-à-vis the original problem, our main theoretical result establishes that the mean taken from the optimal Gaussian policy of the auxiliary problem is exactly the optimal control of the original no-shorting MV problem. In other words, although individual controls sampled from an exploratory policy may not be feasible, they are for the training purpose only. The policy that is ultimately extracted and implemented from that distribution is not only admissible, but indeed is optimal. This underpins theoretically the connection between tractable Gaussian exploration and the original constrained problem.

It should be noted that the above idea was partially inspired by Dai et al. (2025) where a Merton problem is studied. The problem therein does not generate Gaussian exploration if one follows the entropy-regularized formulation of Wang et al. (2020). Instead, the authors propose a different exploratory formulation *without* entropy, which is to randomize controls by simply adding proper zero-mean Gaussian noises to them. They then prove that the mean of their optimal exploratory policy solves the original Merton problem. The problem in Dai et al. (2025) is unconstrained, but here we prove that the key idea can be extended to our constrained problem.

Based on the established theoretical result, we only need to solve the auxiliary exploratory problem for which existing RL theory and algorithms can be adapted to. Indeed, we devise a model-free RL learning algorithm following martingale-based policy evaluation from Jia and Zhou (2022a) and policy gradient from Jia and Zhou (2022b). The algorithm learns the optimal no-shorting portfolio policies from sample paths (i.e. data) without estimation of the investment opportunity set. We also present a numerical example that demonstrates that the algorithm converges and learns the optimal oracle values effectively.

The rest of the paper is organized as follows. In Section 2, we formulate the target mean–variance problem with no-shorting constraints and unknown model parameters and introduce the auxiliary exploratory problem. In Section 3, we carry out a theoretical analysis on the auxiliary problem and establish its relation to the target problem. In Section 4, we develop a model-free RL algorithm for solving the auxiliary (and hence the target) problem. Section 5 presents a numerical example. Section 6 finally concludes.

2 Problem Formulation

In this section, we state the mean–variance problem with no-shorting constraints and unknown parameters (the *target problem*), and then formulate an associated exploratory problem with integral constraints that can be solved in a data-driven manner by an RL algorithm developed in this paper (the *auxiliary problem*). The formulation of the auxiliary problem, while originally rooted in the entropy regularization framework of Wang and Zhou (2020), is directly inspired by Dai et al. (2025) that employs a different way to carry out exploration without involving entropy. This choice of formulation is key to treating the no-shorting constraint.

2.1 Notation

Frequently used notation in this paper includes:

- $M^\intercal$: the transpose of any matrix or vector M;
- $\|M\|$: $\sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix or vector $M = (m_{ij})$;
- $\mathbb{R}^n$: n-dimensional real Euclidean space;
- $\mathbb{R}^n_+$: the subset of $\mathbb{R}^n$ consisting of elements with nonnegative components;
- $\mathcal{N}(\mu, \Sigma)$: the Gaussian density function with mean μ and covariance Σ;
- $\mathcal{N}(\mu, 0)$: the Dirac distribution centered at μ;
- $\mathcal{P}(\mathbb{R}^n)$: the set of density functions on $\mathbb{R}^n$;
- $\mathcal{P}_+(\mathbb{R}^n)$: the subset of $\mathcal{P}(\mathbb{R}^n)$ consisting of distributions with nonnegative mean;
- $\mathbb{S}^n_{++}$: the set of strictly positive definite matrices in $\mathbb{R}^{n\times n}$;
- $a \geq 0$: a componentwise inequality if a is a vector.

Let $T > 0$ be the investment horizon, and $\{W_t,\ 0 \leq t \leq T\}$ be a standard m-dimensional Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_{0\leq t\leq T}, \mathbb{P})$ that satisfies the usual conditions. Given a Hilbert space $\mathcal{H}$ with the norm $\|\cdot\|_{\mathcal{H}}$, define the Banach space

$$\mathcal{L}^2_{\mathcal{F}}(0,T;\mathcal{H}) := \left\{ \varphi(\cdot) \;\middle|\; \varphi \text{ is an } \{\mathcal{F}_t\}_{0\leq t\leq T}\text{-adapted, } \mathcal{H}\text{-valued measurable process} \right.$$
$$\left. \text{and } \mathbb{E}\int_0^T \|\varphi(t)\|^2_{\mathcal{H}}\, dt < \infty \right\}$$

with the norm $\|\varphi\| := \sqrt{\mathbb{E}\int_0^T \|\varphi(t)\|^2_{\mathcal{H}}\, dt}$.

2.2 Mean–Variance Portfolio Selection with No-Shorting Constraints

We consider investing m risky assets whose prices are geometric Brownian motion governed by

$$dS^i_t = S^i_t b^i_t\, dt + S^i_t \sum_{j=1}^m \sigma^{ij}_t\, dW^j_t, \quad t \in [0,T],\ S^i_0 \in \mathbb{R}, \quad i \in \{1,2,\ldots,m\},$$

and one riskless asset whose price follows

$$dS^0_t = S^0_t r_t\, dt, \quad t \in [0,T],\ S^0_0 \in \mathbb{R}.$$

Denote the appreciation rate vector by $b_t := (b^1_t, b^2_t, \ldots, b^m_t)^\intercal$ and the volatility matrix by $\sigma_t := (\sigma^{ij}_t)$. Throughout this paper, we impose the following standing assumptions: 1) the coefficients r_t, b_t, and σ_t are deterministic, measurable, and bounded functions of t on $[0,T]$; and 2) there exists a constant $\delta > 0$ such that $\sigma_t\sigma_t^\intercal - \delta I$ is strictly positive definite for all $t \in [0,T]$.

Let $u(\cdot) = \{u_t,\ 0 \leq t \leq T\}$ be the portfolio (process), where $u_t := (u^1_t, u^2_t, \ldots, u^m_t)^\intercal$ with u^i_t being the *discounted dollar value* in the i-th risky asset at t. This paper studies the case when shorting is disallowed, i.e., $u^i_t \geq 0,\ i \in \{1,2,\ldots,m\},\ t \in [0,T]$. The set of admissible portfolios is defined to be $\mathcal{L}^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)$. For any $u(\cdot) \in \mathcal{L}^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)$, the *discounted wealth process* $x(\cdot) = \{x_t,\ 0 \leq t \leq T\}$ satisfies

$$dx_t = u_t^\intercal \sigma_t(\rho_t\, dt + dW_t), \quad t \in [0,T],\ x_0 \in \mathbb{R}, \tag{1}$$

where $\rho_t := \sigma_t^{-1}(b_t - r_t\mathbf{1})$ is the market price of risk at t.

The mean–variance model under no-shorting constraints is formulated as

$$\min_{u(\cdot)\in\mathcal{L}^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} \mathrm{Var}(x_T), \quad \text{subject to (s.t.)} \quad \mathbb{E}[x_T] = z, \tag{2}$$

where the wealth–portfolio pair $(x(\cdot), u(\cdot))$ satisfies (1), and $z \in \mathbb{R}$ is a pre-specified (discounted) target expected return.

Following Zhou and Li (2000), we introduce a Lagrange multiplier $w \in \mathbb{R}$ for the equality constraint $\mathbb{E}[x_T] = z$. The problem (2) is thus transformed into a standard stochastic optimal control problem with a parameter $w \in \mathbb{R}$:

$$\min_{u(\cdot)\in\mathcal{L}^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J\big(u(\cdot); w\big) := \mathbb{E}[(x_T - w)^2] - (w - z)^2, \quad \text{s. t.} \quad (1). \tag{3}$$

For each fixed $w \in \mathbb{R}$, let $(x^*(\cdot; w), u^*(\cdot; w))$ denote the optimal wealth–portfolio process of (3). Then the correct multiplier w^* can be determined by

$$\mathbb{E}[x^*(T; w^*)] = z. \tag{4}$$

In particular, when $w = w^*$, the problems (2) and (3) have the same optimal portfolio and the same optimal objective value.

The above formulation of the problem is the same as that in Li et al. (2002), except now we do not know the coefficient functions (or the investment opportunity set) (b_t, r_t, σ_t).

2.3 An Auxiliary Problem with Randomized Portfolios

The key feature of this paper is that the model parameters of the problem (2) are unknown. To cope with it we employ the idea of exploratory (randomized) formulation originally proposed by Wang and Zhou (2020) for continuous-time RL. However, directly using the entropy-regularization technique therein runs into the problem of truncated distribution for portfolio randomization due to the no-shorting constraint, as discussed in Introduction. Instead, we follow Dai et al. (2025) to propose an auxiliary problem which is solvable by RL algorithms and, more importantly, whose solution leads directly to the solution to the target problem (2).

In the auxiliary problem, the portfolio $u(\cdot)$ is replaced by a randomized portfolio process $\pi(\cdot)$, which is a distribution (density-function)-valued, progressively measurable stochastic process $\pi(\cdot) := \{\pi_t,\ 0 \le t \le T\}$. Introduce the nonnegative integral constraints on $\pi(\cdot)$:

$$\int_{\mathbb{R}^m} u\,\pi_t(u)\,du \ge 0, \qquad \forall t \in [0, T].$$

In other words, we require that at each time instant t, the distribution π_t must have a nonnegative *mean*. Any such portfolio is called *admissible*. Note that an individual portfolio sampled from an admissible $\pi(\cdot)$ may *violate* the no-shorting constraint.

Recall that $\mathcal{P}_+(\mathbb{R}^m) = \{\pi \in \mathcal{P}(\mathbb{R}^m) \mid \int_{\mathbb{R}^m} u\,\pi(u)\,du \ge 0\}$. Then, the above constraint can also be written as $\pi_t \in \mathcal{P}_+(\mathbb{R}^m)\ \ \forall t \in [0, T]$. Clearly, $\mathcal{P}(\mathbb{R}^m_+) \subset \mathcal{P}_+(\mathbb{R}^m) \subset \mathcal{P}(\mathbb{R}^m)$.

Given a randomized portfolio $\pi(\cdot)$, introduce the exploratory dynamics following Wang et al. (2020):

$$dx_t^\pi = \tilde{b}_t(\pi_t)\,dt + \sum_{j=1}^m \tilde{\sigma}_t^j(\pi_t)\,dW_t^j, \quad t \in [0, T], \tag{5}$$

where $\tilde{b}_t(\cdot)$ and $\tilde{\sigma}_t(\cdot)$ are functions on $\mathcal{P}(\mathbb{R}^m)$ defined by

$$\tilde{b}_t(\pi) := \int_{\mathbb{R}^m} (\sigma_t\rho_t)^\intercal u\,\pi(u)\,du, \qquad \tilde{\sigma}_t^j(\pi) := \left(\int_{\mathbb{R}^m} \Big(\sum_{i=1}^m \sigma_t^{ij} u_i\Big)^2 \pi(u)\,du\right)^{1/2}.$$

We now specify the set of randomized feedback portfolios (*policies*) for our auxiliary problem. A feedback policy $\pi : [0,T]\times\mathbb{R} \to \mathcal{P}_+(\mathbb{R}^m)$ is called *admissible* if the generated process $\{\pi(t, x_t^\pi),\ 0 \le t \le T\}$ is admissible (in particular it satisfies the nonnegative mean constraint), where $\{x_t^\pi,\ 0 \le t \le T\}$ is the solution to (5) with π_t therein substituted with $\pi(t, x_t^\pi)$. Fix a measurable function $\Sigma : [0,T]\times\mathbb{R} \to \mathbb{S}_{++}^m$, and denote by Π the set of admissible *Gaussian* policies π where $\pi(t,x) = \mathcal{N}(a(t,x), \Sigma(t,x))$ for some measurable function $a : [0,T]\times\mathbb{R} \to \mathbb{R}_+^m$.

The proposed auxiliary problem is

$$\min_{\pi\in\Pi} \tilde{J}\big(\pi(\cdot); w\big) := \mathbb{E}[(x_T^\pi - w)^2] - (w-z)^2, \quad \text{s. t.} \quad (5). \tag{6}$$

We highlight some important points on the auxiliary problem (6) along with its relationship with the target problem (3). First, for $\pi \in \Pi$, the distribution π_t at any t has support on the whole space instead of just the positive half plane. Nevertheless, the resulting mean part of the policy $a_t := \int u\,\pi_t(u)\,du \ge 0$ always satisfies the no-shorting constraint by construction. This represents a crucially different way of randomization compared to Wang et al. (2020). Therein, policy randomization is carried out on the action (control) space. Therefore, if the action space is constrained (i.e. truncated), the randomization is on a *truncated* space. By contrast, the randomization introduced here is on the *entire* space even if the action space is constrained. Indeed, we can write $\pi(t,x) = \mathcal{N}(a(t,x), \Sigma(t,x)) = a(t,x) + \mathcal{N}(0, \Sigma(t,x))$ where a is an *admissible* (no-shorting) policy for the target problem, namely, an admissible randomized policy is just to add some proper Gaussian noise to an originally admissible policy. Second, we will show that applying RL we can solve the auxiliary problem (6), without having to know the model coefficients, to obtain an admissible Gaussian policy $\mathcal{N}(a(\cdot,\cdot); \Sigma(\cdot,\cdot))$. Moreover, as one of the main results of this paper, we prove that under some suitable choice of the variance function $\Sigma(\cdot,\cdot)$, the mean $a_t^* := \int u\,\pi_t^*(u)\,du$ of the optimal Gaussian solution to the auxiliary problem, which automatically satisfies the no-shorting constraint, is optimal for the target problem.

3 Analyzing the Auxiliary Problem

In this section, we present an explicit solution of the auxiliary problem (6), and compare it with that of the target problem (3), in terms of the model parameters ρ_t and σ_t. A key is to exploit the freedom of choosing the variance function $\Sigma(\cdot,\cdot)$ to relate the two problems. These results establish the theoretical foundation for the model-free algorithm to be introduced in the next section.

Define the optimal value function of (6) by

$$\tilde{V}(t,x;w) := \inf_{\pi\in\Pi} \mathbb{E}[(x_T^\pi - w)^2 - (w-z)^2 \mid x_t^\pi = x].$$

The associated Hamiltonian is

$$\begin{aligned} H(t,x,\pi) &= \frac{\partial}{\partial x}\tilde{V}(t,x;w)\tilde{b}_t(\pi) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\tilde{V}(t,x;w)\|\tilde{\sigma}_t(\pi)\|^2 \\ &= \int_{\mathbb{R}^m}\left(\frac{\partial}{\partial x}\tilde{V}(t,x;w)u^\intercal\sigma_t\rho_t + \frac{1}{2}\frac{\partial^2}{\partial x^2}\tilde{V}(t,x;w)u^\intercal\sigma_t\sigma_t^\intercal u\right)\pi(u)\,du, \end{aligned}$$

where $(t,x,\pi) \in [0,T]\times\mathbb{R}\times\mathcal{P}_+(\mathbb{R}^m)$. Hence, the HJB equation for (6) is

$$\begin{cases} \dfrac{\partial}{\partial t}v(t,x) + \inf\limits_{\pi\in\mathcal{P}_+(\mathbb{R}^m)} \displaystyle\int\left(\frac{\partial}{\partial x}v(t,x)\,u^\intercal\sigma_t\rho_t + \frac{1}{2}\frac{\partial^2}{\partial x^2}v(t,x)u^\intercal\sigma_t\sigma_t^\intercal u\right)\pi(u)\,du = 0, \\ v(T,x) = (x-w)^2 - (w-z)^2. \end{cases} \tag{7}$$

However, for each (t,x) the randomized control π is of the form $\pi = \mathcal{N}(a, \Sigma(t,x))$ with the mean vector $a \in \mathbb{R}_+^m$ and a fixed, chosen $\Sigma(t,x)$. Substituting this form into (7) we can further rewrite the equation

as

$$\begin{cases} \dfrac{\partial}{\partial t}v(t,x) + \inf\limits_{a\in\mathbb{R}^m_+}\left(\dfrac{\partial}{\partial x}v(t,x)\, a^\intercal\sigma_t\rho_t + \dfrac{1}{2}\dfrac{\partial^2}{\partial x^2}v(t,x)a^\intercal\sigma_t\sigma_t^\intercal a\right) + \dfrac{1}{2}\dfrac{\partial^2}{\partial x^2}v(t,x)\,\mathrm{tr}[\Sigma(t,x)\sigma_t\sigma_t^\intercal] = 0, \\ v(T,x) = (x-w)^2 - (w-z)^2. \end{cases} \tag{8}$$

The solution of (8) depends on the specific choice of the variance function $\Sigma(\cdot,\cdot)$. Note that when $\Sigma(\cdot,\cdot)\equiv 0$, (8) reduces to the HJB equation for the target problem; see Li et al. (2002).

Define $\mu_t^{\Gamma_1},\ \mu_t^{\Gamma_2},\ P_t^{\Gamma_1},\ P_t^{\Gamma_2}$ by

$$\begin{aligned} \mu_t^{\Gamma_1} &:= \operatorname*{argmin}_{\mu\in\mathbb{R}^m_+}\left[\frac{1}{2}\mu^\intercal\sigma_t\sigma_t^\intercal\mu - (\sigma_t\rho_t)^\intercal\mu\right], \\ \mu_t^{\Gamma_2} &:= \operatorname*{argmin}_{\mu\in\mathbb{R}^m_+}\left[\frac{1}{2}\mu^\intercal\sigma_t\sigma_t^\intercal\mu + (\sigma_t\rho_t)^\intercal\mu\right], \\ P_t^{\Gamma_1} &:= \exp\left(\int_t^T [(\mu_s^{\Gamma_1})^\intercal\sigma_s\sigma_s^\intercal(\mu_s^{\Gamma_1}) - 2(\sigma_s\rho_s)^\intercal\mu_s^{\Gamma_1}]\,ds\right), \\ P_t^{\Gamma_2} &:= \exp\left(\int_t^T [(\mu_s^{\Gamma_2})^\intercal\sigma_s\sigma_s^\intercal(\mu_s^{\Gamma_2}) + 2(\sigma_s\rho_s)^\intercal\mu_s^{\Gamma_2}]\,ds\right). \end{aligned} \tag{9}$$

Clearly, we have $\mu_t^{\Gamma_1}\geq 0,\ \mu_t^{\Gamma_2}\geq 0,\ 0<P_t^{\Gamma_1}\leq 1$, and $0<P_t^{\Gamma_2}\leq 1$ for all t.

We now specify the following *state-dependent* variance function for exploration:

$$\Sigma(t,x) = (x-w)^2\Psi_t,$$

where $\Psi:[0,T]\to\mathbb{S}^m_{++}$ is any given matrix-valued measurable function.

Define

$$\tilde{P}_t^{\Gamma_1} := P_t^{\Gamma_1}\exp\left(\int_t^T \mathrm{tr}[\Psi_s\sigma_s\sigma_s^\intercal]\,ds\right), \qquad \tilde{P}_t^{\Gamma_2} := P_t^{\Gamma_2}\exp\left(\int_t^T \mathrm{tr}[\Psi_s\sigma_s\sigma_s^\intercal]\,ds\right). \tag{10}$$

We are going to show that

$$\tilde{V}^{\Gamma_1}(t,x;w) := \tilde{P}_t^{\Gamma_1}(x-w)^2 - (w-z)^2$$

solves the HJB equation (8) in the region $\Gamma_1 := \{(t,x)\mid x-w<0\}$, and

$$\tilde{V}^{\Gamma_2}(t,x;w) := \tilde{P}_t^{\Gamma_2}(x-w)^2 - (w-z)^2$$

solves the HJB equation (8) in the region $\Gamma_2 := \{(t,x)\mid x-w>0\}$ respectively.

To this end, we first need the following scaling lemma.

Lemma 1. *Let $\Sigma\in\mathbb{R}^{m\times m}$, $p\in\mathbb{R}^m$ and $s>0$. Assume that Σ is strictly positive definite. Then,*

$$\min_{a\in\mathbb{R}^m_+}[\frac{1}{2}a^\intercal\Sigma a + sp^\intercal a] = s^2\cdot\min_{\mu\in\mathbb{R}^m_+}[\frac{1}{2}\mu^\intercal\Sigma\mu + p^\intercal\mu],$$

and

$$\operatorname*{argmin}_{a\in\mathbb{R}^m_+}[\frac{1}{2}a^\intercal\Sigma a + sp^\intercal a] = s\cdot\operatorname*{argmin}_{\mu\in\mathbb{R}^m_+}[\frac{1}{2}\mu^\intercal\Sigma\mu + p^\intercal\mu].$$

Proof. We apply the standard Karush–Kuhn–Tucker (KKT) condition. Denote $f(a;s) = \frac{1}{2}a^\intercal\Sigma a + sp^\intercal a$. Since Σ is strictly positive definite, $f(\cdot;s)$ has a unique minimizer on $\mathbb{R}^m_+$.

Consider the convex minimization problem with linear inequality constraints

$$\min_{a\in\mathbb{R}^m} \quad f(a;s), \quad \text{s.t.} \quad a\geq 0. \tag{11}$$

Recall that a pair $(a, \lambda) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ is said to be a KKT pair of (11) if

$$\lambda^\intercal a = 0, \qquad \lambda = \Sigma a + sp. \tag{12}$$

By standard results in convex analysis (see, e.g., Boyd and Vandenberghe 2023, pp. 224–226), a pair (a, λ) is a KKT pair of (11) if and only of a is its minimizer.

Fix $s > 0$. Let $(a_s, \lambda_s) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ be a KKT pair of (11). Then,

$$\left(\frac{\lambda_s}{s}\right)^\intercal \left(\frac{a_s}{s}\right) = 0, \qquad \frac{\lambda_s}{s} = \Sigma \frac{a_s}{s} + p.$$

This shows that $(\frac{a_s}{s}, \frac{\lambda_s}{s}) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ is a KKT pair of (11) at $s = 1$. Hence,

$$a_s = \operatorname*{argmin}_{a \in \mathbb{R}^m_+} f(a; s), \qquad \frac{a_s}{s} = \operatorname*{argmin}_{a \in \mathbb{R}^m_+} f(a; 1).$$

This proves the second claim of the lemma. The first claim is obtained directly from the fact that $f(a_s, s) = s^2 f(\frac{a_s}{s}, 1)$. □

Now consider the region Γ_1 where $x - w < 0$. Substituting $\frac{\partial^2}{\partial x^2}\tilde{V}^{\Gamma_1}(t, x; w) = 2\tilde{P}_t^{\Gamma_1} > 0$ and $\frac{\partial}{\partial x}\tilde{V}^{\Gamma_1}(t, x; w) = 2\tilde{P}_t^{\Gamma_1}(x - w)$ into the left-hand side of (8) yields

$$\begin{aligned}
\text{LHS} &= \dot{\tilde{P}}_t^{\Gamma_1}(x-w)^2 + 2\tilde{P}_t^{\Gamma_1} \inf_{a \in \mathbb{R}^m_+} \left(\frac{1}{2} a^\intercal \sigma_t \sigma_t^\intercal a + (x-w) a^\intercal \sigma_t \rho_t \right) + \tilde{P}_t^{\Gamma_1}(x-w)^2 \operatorname{tr}[\Psi_t \sigma_t \sigma_t^\intercal] \\
&= \dot{\tilde{P}}_t^{\Gamma_1}(x-w)^2 + 2\tilde{P}_t^{\Gamma_1} \inf_{a \in \mathbb{R}^m_+} \left(\frac{1}{2} a^\intercal \sigma_t \sigma_t^\intercal a - |x-w| (\sigma_t \rho_t)^\intercal a \right) + \tilde{P}_t^{\Gamma_1}(x-w)^2 \operatorname{tr}[\Psi_t \sigma_t \sigma_t^\intercal].
\end{aligned}$$

Applying Lemma 1 with the scaling parameter $|x - w|$, we obtain

$$\text{LHS} = \left[\dot{\tilde{P}}_t^{\Gamma_1} + \tilde{P}_t^{\Gamma_1} \left((\mu_t^{\Gamma_1})^\intercal \sigma_t \sigma_t^\intercal \mu_t^{\Gamma_1} - 2(\sigma_t \rho_t)^\intercal \mu_t^{\Gamma_1} + \operatorname{tr}\{\Psi_t \sigma_t \sigma_t^\intercal\} \right) \right] (x-w)^2 = 0.$$

This verifies the solution in Γ_1. The case for Γ_2 is completely analogous and omitted here.

Next, we are to show that the optimal value function of the auxiliary problem is

$$\tilde{V}(t, x; w) = \begin{cases} \tilde{P}_t^{\Gamma_1}(x-w)^2 - (w-z)^2, & \text{if } x - w < 0, \\ \tilde{P}_t^{\Gamma_2}(x-w)^2 - (w-z)^2, & \text{if } x - w > 0, \\ -(w-z)^2, & \text{if } x - w = 0. \end{cases} \tag{13}$$

In fact, this function is continuously differentiable on $\Gamma_3 := \{(t, x) \mid x = w\}$ because

$$\begin{aligned}
\lim_{x \to w^-} \tilde{V}(t, x; w) &= \lim_{x \to w^+} \tilde{V}(t, x; w) = -(w-z)^2, \\
\lim_{x \to w^-} \frac{\partial}{\partial x} \tilde{V}(t, x; w) &= \lim_{x \to w^+} \frac{\partial}{\partial x} \tilde{V}(t, x; w) = 0.
\end{aligned} \tag{14}$$

It also clearly satisfies the boundary condition of the HJB equation (8) as $\tilde{P}_T^{\Gamma_1} = \tilde{P}_T^{\Gamma_2} = 1$. Furthermore, (13) satisfies (8) in the viscosity sense; see the proof of Theorem 1 below and Appendix A for the precise definition of viscosity solutions. Thus the standard uniqueness result of viscosity solution establishes that (13) is the value function of the auxiliary problem.

It follows that we can obtain the optimal mean strategy a^* by minimizing the Hamiltonian

$$\begin{aligned} a^*(t,x;w) &= \operatorname*{argmin}_{a\in\mathbb{R}^m_+}\left(\frac{\partial}{\partial x}\tilde{V}(t,x;w)a^\intercal\sigma_t\rho_t + \frac{1}{2}\frac{\partial^2}{\partial x^2}\tilde{V}(t,x;w)a^\intercal\sigma_t\sigma_t^\intercal a\right) \\ &= \operatorname*{argmin}_{a\in\mathbb{R}^m_+}\left(\frac{1}{2}a^\intercal\sigma_t\sigma_t^\intercal a + \frac{\frac{\partial}{\partial x}\tilde{V}(t,x;w)}{\frac{\partial^2}{\partial x^2}\tilde{V}(t,x;w)}a^\intercal\sigma_t\rho_t\right) \\ &= \operatorname*{argmin}_{a\in\mathbb{R}^m_+}\left(\frac{1}{2}a^\intercal\sigma_t\sigma_t^\intercal a + (x-w)a^\intercal\sigma_t\rho_t\right) \\ &= \begin{cases} -(x-w)\mu_t^{\Gamma_1}, & \text{if } x-w<0, \\ (x-w)\mu_t^{\Gamma_2}, & \text{if } x-w\geq 0. \end{cases} \end{aligned}$$

The obtained optimal mean strategy a^* turns out to be identical to the theoretically optimal strategy u^* of the target problem, derived in Li et al. (2002):

$$u^*(t,x;w) = \begin{cases} -(x-w)\mu_t^{\Gamma_1}, & \text{if } x-w<0, \\ (x-w)\mu_t^{\Gamma_2}, & \text{if } x-w\geq 0. \end{cases} \tag{15}$$

We arrive at the following theorem.

Theorem 1. *Let $\Psi:[0,T]\to\mathbb{S}^m_{++}$ be a matrix-valued measurable function, and Π be the policy space defined in (6) with the variance function set to be $\Sigma(t,x)=(x-w)^2\Psi_t$. Then, the HJB equation (8) has a unique viscosity solution (13) which is also the optimal value function of the auxiliary problem (6). Moreover, the optimal strategy of (6) is*

$$\pi^*(t,x;w) = \begin{cases} \mathcal{N}\big(-(x-w)\mu_t^{\Gamma_1},(x-w)^2\Psi_t\big), & \text{if } x-w<0, \\ \mathcal{N}\big((x-w)\mu_t^{\Gamma_2},(x-w)^2\Psi_t\big), & \text{if } x-w\geq 0, \end{cases} \tag{16}$$

where $\mu_t^{\Gamma_1}$ and $\mu_t^{\Gamma_2}$ are given by (9). Furthermore, the mean strategy $a^*(t,x;w):=\int_{\mathbb{R}^m}u\pi^*(u\mid t,x;w)\,du$ is optimal to the target problem (3).

Proof. Let $\tilde{V}$ be the piecewise quadratic function defined by (13), which clearly satisfies the terminal condition.

1. *Verification in the interior regions Γ_i $(i=1,2)$.* When $(t,x)\in\Gamma_i$, we have

$$\tilde{V}(t,x;w) = \tilde{P}_t^{\Gamma_i}(x-w)^2 - (w-z)^2.$$

Thus,

$$\frac{\partial\tilde{V}}{\partial t} = \dot{\tilde{P}}_t^{\Gamma_i}(x-w)^2, \quad \frac{\partial\tilde{V}}{\partial x} = 2\tilde{P}_t^{\Gamma_i}(x-w), \quad \frac{\partial^2\tilde{V}}{\partial x^2} = 2\tilde{P}_t^{\Gamma_i}.$$

Substituting these into the left-hand side of (8) yields

$$\begin{aligned} \text{LHS} &= \dot{\tilde{P}}_t^{\Gamma_i}(x-w)^2 + \inf_{a\in\mathbb{R}^m_+}\left\{2\tilde{P}_t^{\Gamma_i}(x-w)a^\intercal\sigma_t\rho_t + \tilde{P}_t^{\Gamma_i}a^\intercal\sigma_t\sigma_t^\intercal a\right\} + \tilde{P}_t^{\Gamma_i}(x-w)^2\mathrm{tr}\{\Psi_t\sigma_t\sigma_t^\intercal\} \\ &= (x-w)^2[\dot{\tilde{P}}_t^{\Gamma_i} + \tilde{P}_t^{\Gamma_i}\mathrm{tr}\{\Psi_t\sigma_t\sigma_t^\intercal\}] + 2\tilde{P}_t^{\Gamma_i}\inf_{a\in\mathbb{R}^m_+}\left\{(x-w)a^\intercal\sigma_t\rho_t + \frac{1}{2}a^\intercal\sigma_t\sigma_t^\intercal a\right\}. \end{aligned}$$

Since $\tilde{P}_t^{\Gamma_i} = P_t^{\Gamma_i}\exp\big(\int_t^T\mathrm{tr}\{\Psi_s\sigma_s\sigma_s^\intercal\}\,ds\big)$, we have

$$\dot{\tilde{P}}_t^{\Gamma_i} + \tilde{P}_t^{\Gamma_i}\mathrm{tr}\{\Psi_t\sigma_t\sigma_t^\intercal\} = \dot{P}_t^{\Gamma_i}\exp\left(\int_t^T\mathrm{tr}\{\Psi_s\sigma_s\sigma_s^\intercal\}\,ds\right) = \dot{P}_t^{\Gamma_i}\frac{\tilde{P}_t^{\Gamma_i}}{P_t^{\Gamma_i}}.$$

Therefore,

$$\text{LHS} = \frac{\tilde{P}_t^{\Gamma_i}}{P_t^{\Gamma_i}}\left[\dot{P}_t^{\Gamma_i}(x-w)^2 + 2P_t^{\Gamma_i}\inf_{a\in\mathbb{R}^m_+}\left\{(x-w)a^\intercal\sigma_t\rho_t + \frac{1}{2}a^\intercal\sigma_t\sigma_t^\intercal a\right\}\right].$$

The term inside the brackets is exactly the HJB equation for the target problem and thus equals zero. This verifies that $\tilde{V}$ satisfies the HJB equation in the regions Γ_i $(i = 1, 2)$.

2. Verification on the boundary Γ_3. (14) ensures that both $\tilde{V}$ and $\frac{\partial}{\partial x}\tilde{V}$ at points on Γ_3. Moreover, we have

$$\lim_{x \to w^-} \frac{\partial^2}{\partial x^2} \tilde{V}(t, x; w) = \tilde{P}_t^{\Gamma_1},$$
$$\lim_{x \to w^+} \frac{\partial^2}{\partial x^2} \tilde{V}(t, x; w) = \tilde{P}_t^{\Gamma_2},$$

implying that for all $(t, x) \in \Gamma_3$,

$$D_{t,x}^{1,2,+} \tilde{V}(t, x) = \{0\} \times \{0\} \times [\tilde{P}_t^{\Gamma_2}, \infty),$$
$$D_{t,x}^{1,2,-} \tilde{V}(t, x) = \{0\} \times \{0\} \times (-\infty, \tilde{P}_t^{\Gamma_1}].$$

Therefore, for any $(t, x) \in \Gamma_3$ and $(q, p, P) \in D_{t,x}^{1,2,+} \tilde{V}(t, x)$,

$$q + \inf_{a \geq 0} G(t, x, a, p, P) \geq 0 + \inf_{a \geq 0} G(t, x, a, 0, \tilde{P}_t^{\Gamma_2}) = 0,$$

and for any $(t, x) \in \Gamma_3$ and $(q, p, P) \in D_{t,x}^{1,2,-} \tilde{V}(t, x)$,

$$q + \inf_{a \geq 0} G(t, x, a, p, P) \leq 0 + \inf_{a \geq 0} G(t, x, a, 0, \tilde{P}_t^{\Gamma_1}) = 0.$$

This establishes that $\tilde{V}$ is a viscosity solution of (8). The uniqueness of the viscosity solution is standard; see e.g. Yong and Zhou (1999). □

Several remarks are in order. First off, in the next section we will show that the randomized auxiliary problem (6) with unknown parameters can be solved algorithmically using the recently developed continuous-time RL theory. Theorem 1, on the other hand, indicates that the mean of the optimal Gaussian policy recovers the solution to the original target problem (3). Crucially, because the mean of any admissible Gaussian policy is nonnegative by construction, it satisfies the no-shorting constraint of (3). Second, the function Ψ_t does not affect the optimal mean policy; it controls only the exploration level via the variance of the underlying Gaussian policy. Finally, both the mean and the variance of $\pi^*(t, x; w)$ vanish as x approaches w, and at $x = w$ the policy degenerates to Dirac at 0 (i.e. completely no risky allocation) and, as a result, the (discounted) wealth stays put at $x = w$. In other words, $x = w$ is a "dividing boundary" of the state space under π^*, preventing any admissible wealth trajectory from crossing between Γ_1 and Γ_2.

4 Solving Portfolio Selection with Unknown Dynamics

In this section, we develop a model-free algorithm to learn the optimal policy of the target portfolio selection problem (3) without knowing the investment opportunity set. Based on the theoretical results presented in the previous section, it is equivalent to searching for the solution of the auxiliary, exploratory problem (6). To do so we apply a trajectory-based (i.e. data-driven) policy gradient estimator and update our policy within the feasible set via the proximal gradient descent method.

4.1 Parameterization and Policy Evaluation

We take $\Sigma(t, x) = (x - w)^2 \Psi_0$ where $\Psi_0 \in \mathbb{S}_{++}^n$ is a fixed matrix. In implementation, one can choose $\Psi_0 = \lambda I_m$ where λ is a tuning hyperparameter controlling the exploration strength. Consider the following parameterized policy:

$$\pi^\phi(t, x; w) = \mathcal{N}\big(u^\phi(t, x; w), (x - w)^2 \Psi_0\big), \tag{17}$$

where $u^\phi(t,x;w)$ is a parameterized function that takes value in $\mathbb{R}^m_+$.

In general, one can use feedforward neural networks to represent u^ϕ and set the last layer to a ReLU operation to ensure the nonnegativity (see, e.g., Bishop and Bishop 2024). In this paper, however, we use a piecewise linear function to represent u^ϕ, inspired by the theoretical structure of the optimal control (15). Specifically, the parameterized function is set to be

$$u^\phi(t,x;w) = \begin{cases} -(x-w)\phi_t^{(1)}, & \text{if } x-w<0, \\ (x-w)\phi_t^{(2)}, & \text{if } x-w\geq 0. \end{cases} \tag{18}$$

For numerical implementation, the time interval $[0,T]$ is discretized into a finite grid $\{i\frac{T}{N},\quad i=0,1,2,\dots N\}$. Let the values of the functions $\phi^{(1)}$ and $\phi^{(2)}$ at the time instant $t_i = i\frac{T}{N}$ be $\phi_i^{(1)}\in\mathbb{R}^m_+$ and $\phi_i^{(2)}\in\mathbb{R}^m_+$ respectively. Hence, there are $2N+2$ vectors to be optimized.[1] With a slight abuse of notation, henceforth we use ϕ to denote both the function $\phi := (\phi^{(1)},\phi^{(2)})$ and the sequence of vectors $\phi := \{\phi_i^{(1)},\phi_i^{(2)}\}_{i=0}^N$.

The value function (i.e. cost function) of π^ϕ is

$$V^\phi(t,x;w) = \mathbb{E}[(x_T^\phi - w)^2 - (w-z)^2 \mid x_t^\phi = x] \tag{19}$$

where x^ϕ is the solution of the exploratory dynamics (5) under π^ϕ. The piecewise linear policy (18) suggests an explicit way to parameterize V^ϕ. Indeed, by the Feynman–Kac formula, the right hand side of (19) satisfies the PDE

$$\begin{cases} \dfrac{\partial}{\partial t}v(t,x) + \dfrac{\partial}{\partial x}v(t,x)\,(\sigma_t\rho_t)^\intercal u^\phi(t,x;w) + \dfrac{1}{2}\dfrac{\partial^2}{\partial x^2}v(t,x)(u^\phi(t,x;w))^\intercal\sigma_t\sigma_t^\intercal u^\phi(t,x;w) \\ \qquad\qquad + \dfrac{1}{2}\dfrac{\partial^2}{\partial x^2}v(t,x)\,\mathrm{tr}[\Psi_0\sigma_t\sigma_t^\intercal](x-w)^2 = 0, \\ v(T,x) = (x-w)^2 - (w-z)^2. \end{cases}$$

Solving it yields

$$V^\phi(t,x;w) = \begin{cases} e^{\theta_t^{\Gamma_1}}(x-w)^2 - (w-z)^2, & \text{if } x-w<0, \\ e^{\theta_t^{\Gamma_2}}(x-w)^2 - (w-z)^2, & \text{if } x-w>0, \\ -(w-z)^2, & \text{if } x-w=0, \end{cases} \tag{20}$$

where $\theta_t^{\Gamma_1}$ and $\theta_t^{\Gamma_2}$ are given, theoretically, by

$$\begin{aligned} \theta_t^{\Gamma_1} &:= \int_t^T \left((\phi_s^{(1)})^\intercal\sigma_s\sigma_s^\intercal(\phi_s^{(1)}) - 2(\sigma_s\rho_s)^\intercal\phi_s^{(1)} + \mathrm{tr}[\Psi_0\sigma_s\sigma_s^\intercal] \right) ds, \\ \theta_t^{\Gamma_2} &:= \int_t^T \left((\phi_s^{(2)})^\intercal\sigma_s\sigma_s^\intercal(\phi_s^{(2)}) + 2(\sigma_s\rho_s)^\intercal\phi_s^{(2)} + \mathrm{tr}[\Psi_0\sigma_s\sigma_s^\intercal] \right) ds. \end{aligned} \tag{21}$$

We do not know the functions θ^{Γ_1} and θ^{Γ_2} due to the unknown model parameters; so we parameterize them. Following the parameterization used for $\phi^{(1)}$ and $\phi^{(2)}$, we discretize the trajectories $\{\theta_t^{\Gamma_1},\theta_t^{\Gamma_2}\}_{0\leq t\leq T}$ using $2N+2$ scalars $\{\theta_i^{(1)},\theta_i^{(2)}\}_{i=0}^N$. This is equivalent to the parameterization of the value function in the form

$$V^\theta(t,x;w) = \begin{cases} e^{\theta_t^{(1)}}(x-w)^2 - (w-z)^2, & \text{if } x-w<0, \\ e^{\theta_t^{(2)}}(x-w)^2 - (w-z)^2, & \text{if } x-w>0, \\ -(w-z)^2, & \text{if } x-w=0. \end{cases} \tag{22}$$

As with ϕ, we use θ to denote interchangeably both the function $\theta := (\theta^{(1)},\theta^{(2)})$ and the sequence of vectors $\theta := \{\theta_i^{(1)},\theta_i^{(2)}\}_{i=0}^N$. We now apply the martingale loss method proposed in Jia and Zhou (2022a)

[1] If the portfolio selection problem is time invariant, then they reduce to two vectors, as the coefficients are identical for all t_i.

to perform the policy evaluation, in order to learn $(\theta^{(1)}, \theta^{(2)})$ or equivalently the vectors $\{\theta_i^{(1)}, \theta_i^{(2)}\}_{i=0}^N$ under a given π^ϕ. Let $(x(\cdot), u(\cdot))$ be the wealth–portfolio pair of (1) while $u(\cdot)$ is sampled from the randomized policy $\pi^\phi(\cdot)$. The martingale loss is defined by

$$L(\theta) := \mathbb{E}\int_0^T \left(V^\theta(t, x_t; w) - [(x_T - w)^2 - (w - z)^2]\right)^2 dt.$$

We use the mini-batch stochastic gradient descent (SGD) method to optimize the martingale loss. Discretizing $(x(\cdot), u(\cdot))$ into $\{(x_i, u_i)\}_{i=0}^N$ on the given time grid, we can compute the gradient by

$$\begin{aligned}
\frac{\partial}{\partial \theta_i^{(1)}} L(\theta) &= \begin{cases} \mathbb{E}\left[2\Big(V^\theta(t_i, x_i; w) - \big[(x_N - w)^2 - (w - z)^2\big]\Big) \cdot e^{\theta_i^{(1)}} (x_i - w)^2 \frac{T}{N}\right], & \text{if } x_i - w < 0, \\ 0, & \text{if } x_i - w \geq 0, \end{cases} \\
\frac{\partial}{\partial \theta_i^{(2)}} L(\theta) &= \begin{cases} 0, & \text{if } x_i - w < 0, \\ \mathbb{E}\left[2\Big(V^\theta(t_i, x_i; w) - \big[(x_N - w)^2 - (w - z)^2\big]\Big) \cdot e^{\theta_i^{(2)}} (x_i - w)^2 \frac{T}{N}\right], & \text{if } x_i - w \geq 0. \end{cases}
\end{aligned} \tag{23}$$

As discussed earlier, the policy parameter ϕ is identified as vectors $\{(\phi_i^{(1)}, \phi_i^{(2)})\}_{i=0}^N$ for (18), and the value function parameter θ as vectors $\{(\theta_i^{(1)}, \theta_i^{(2)})\}_{i=0}^N$ for (22). In our implementation, $\phi_N^{(1)}$ and $\phi_N^{(2)}$ are fixed to be 0 as they are never used (the system exits at the terminal time and the terminal control will not be applied). In addition, $\theta_N^{(1)}$ and $\theta_N^{(2)}$ are fixed to be 0 to satisfy the terminal condition $V^\theta(T, x) = (x - w)^2 - (w - z)^2$.

4.2 Policy Gradient

Policy gradient is a popular approach in model-free RL. It uses the learned value function to construct an estimator of its gradient with respect to the policy parameter, providing the direction of moving to the next policy iterate. There are many such estimators proposed, with different trade-offs in terms of efficiency, bias, and variance. In this paper, we use the gradient estimator of Jia and Zhou (2022b) developed for continuous-time RL and hence well suited for the setting of this paper.

Lemma 2. *Let π^ϕ be a parameterized randomized policy, and let*

$$V^\phi(t, x) := \mathbb{E}[g(x_T) \mid x_t = x]$$

be the corresponding value function, where $(x(\cdot), u(\cdot))$ denotes the wealth–portfolio trajectory generated by π^ϕ. Then the policy gradient admits the representation

$$\frac{\partial}{\partial \phi} V^\phi(t, x) = \mathbb{E}\left[\int_t^T \left(\frac{\partial}{\partial \phi} \log \pi^\phi(u_s \mid s, x_s)\right) dV^\phi(s, x_s) \,\middle|\, x_t = x\right]. \tag{24}$$

Proof. The result follows directly from Theorem 5 of Jia and Zhou (2022b). □

Applying Lemma 2 to the exploratory objective $\tilde{J}(\pi^\phi(\cdot); w)$ yields

$$\frac{\partial}{\partial \phi} \tilde{J}(\pi^\phi(\cdot); w) = \frac{\partial}{\partial \phi} V^\phi(0, x; w) = \mathbb{E}\int_0^T \left(\frac{\partial}{\partial \phi} \log \pi^\phi(u_t \mid t, x_t)\right) dV^\phi(t, x_t; w). \tag{25}$$

Recall that π^ϕ is constructed by (17) and (18), whereas its parameter ϕ is taken as $\{(\phi_i^{(1)}, \phi_i^{(2)})\}_{i=0}^N$. Given the discretized wealth–portfolio trajectory $\{(x_i, u_i)\}_{i=0}^N$, the policy gradient (24) in each com-

ponent is

$$\frac{\partial}{\partial \phi_i^{(1)}} \tilde{J}(\pi^\phi(\cdot); w) = \begin{cases} \mathbb{E}\left[-\Psi_0^{-1}\left(\frac{u_i}{x_i - w} + \phi_i^{(1)}\right)\left(V^\phi(t_{i+1}, x_{i+1}; w) - V^\phi(t_i, x_i; w)\right)\right], & \text{if } x_i - w < 0, \\ 0, & \text{if } x_i - w \geq 0, \end{cases}$$

$$\frac{\partial}{\partial \phi_i^{(2)}} \tilde{J}(\pi^\phi(\cdot); w) = \begin{cases} 0, & \text{if } x_i - w < 0, \\ \mathbb{E}\left[-\Psi_0^{-1}\left(\frac{u_i}{x_i - w} + \phi_i^{(2)}\right)\left(V^\phi(t_{i+1}, x_{i+1}; w) - V^\phi(t_i, x_i; w)\right)\right], & \text{if } x_i - w \geq 0. \end{cases} \tag{26}$$

To satisfy the nonnegativity constraint on the mean of π^ϕ, we ensure $\phi_i^{(1)} \geq 0$ and $\phi_i^{(2)} \geq 0$ by projecting them onto $\mathbb{R}_+^m$ after each gradient descent step. This is known as the proximal gradient method for constrained optimization problems (Beck, 2017). In our case, the proximal gradient method is guaranteed to converge because $V^\phi(0, x; w)$ is convex in ϕ; see (20)–(21).

Finally, the update rule of the multiplier w is based on applying the standard stochastic approximation method to the equation (4), where $\mathbb{E}[x^*(T; w)]$ is estimated as the empirical mean over wealth processes under portfolios sampled from π^ϕ.

We summarize the overall procedure as Algorithm 1.

Algorithm 1 Model-Free Policy Gradient for the Mean–Variance Problem

Require: Initial parameters $\phi,\ \theta,\ w$; learning rates $\alpha_\phi,\ \alpha_\theta,\ \alpha_w$; batch size M; iteration counts $n_\phi,\ n_\theta, n_w$, and n_{warmup}.

for $\ell = 1$ **to** n_{warmup} **do**
  Sample M trajectories by executing π^ϕ, organized as $\mathcal{D}_\ell := \{(x_i^{(j)}, u_i^{(j)}),\ i = 0, 1, \ldots, N\}_{j=1}^M$.
  Estimate the gradient $g_\theta \leftarrow \partial_\theta L(\theta)$ using $\mathcal{D}_\ell$ via (23) and update

$$\theta \leftarrow \theta - \alpha_\theta g_\theta.$$

end for
Initialize an empty data buffer $\mathcal{X}$.
for $k = 1$ **to** n_ϕ **do**
  Initialize an empty data buffer $\mathcal{D}$.
  for $\ell = 1$ **to** n_θ **do**
    Sample M trajectories by executing π^ϕ, organized as $\mathcal{D}_\ell := \{(x_i^{(j)}, u_i^{(j)}),\ i = 0, 1, \ldots, N\}_{j=1}^M$.
    Estimate the gradient $g_\theta \leftarrow \partial_\theta L(\theta)$ using $\mathcal{D}_\ell$ via (23) and update

$$\theta \leftarrow \theta - \alpha_\theta g_\theta.$$

    Append $\mathcal{D}_\ell$ to $\mathcal{D}$.
  end for
  Estimate the policy gradient $g_\phi \leftarrow \partial_\phi \tilde{J}(\pi^\phi)$ using $\mathcal{D}$ via (26) and update

$$\phi \leftarrow \max(0, \phi - \alpha_\phi g_\phi).$$

  Compute the empirical mean of x_T in $\mathcal{D}$ and append it to $\mathcal{X}$.
  if $(k \bmod n_w) = 0$ **then**
    Compute $\bar{x}_T$ as the average of the n_w most recent entries in $\mathcal{X}$ and update

$$w \leftarrow w - \alpha_w(\bar{x}_T - z).$$

  end if
end for
return Updated policy π^ϕ and updated multiplier w.

5 A Numerical Example

We test Algorithm 1 on a time-invariant example of the problem (2) under no-shorting constraints, and then compare the results with the theoretical (ground truth) solution obtained from (9) and (15).

Problem data and theoretical solution. We set

$$b - r\mathbf{1} = \begin{bmatrix} 0.06 \\ 0.12 \\ 0.16 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 0.20 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.35 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \sqrt{3}/2 & 0 \\ \frac{1}{2} & \sqrt{3}/6 & \sqrt{6}/3 \end{bmatrix}.$$

Note here $\sigma = \text{diag}\{\nu\}\Lambda$, where ν is the asset volatilities and Λ is the Cholesky factor of the correlation matrix. Moreover, set the investment horizon $T = 1$ and the initial wealth $x_0 = 1$.

In our experiment, we take the exploration variance parameter $\Psi_0 = \frac{1}{5} I_3$, the time discretization step size $\frac{T}{N} = 0.01$, and the batch size $M = 512$ for each simulation.

The theoretical solution (15) of the no-shorting MV problem (2) in this case specializes to

$$u^*(t,x) = \begin{cases} -\phi^*(x - w^*), & \text{if } x < w^*, \\ 0, & \text{if } x \geq w^*, \end{cases}$$

where the pair (ϕ^*, w^*) can be computed from solving the following problem

$$\max_{w\in\mathbb{R}} \min_{\phi\in\mathbb{R}^3_+} J(\phi, w) := e^{[\phi^\intercal \sigma\sigma^\intercal \phi - 2(\sigma\rho)^\intercal \phi]T}(x_0 - w)^2 - (w - z)^2.$$

Note ϕ^* is exactly the coefficient $\mu_t^{\Gamma_1}$ defined in (9), which is time-independent in this example. Moreover, $\mu_t^{\Gamma_2} \equiv 0$ as $\sigma_t\rho_t = b_t - r_t\mathbf{1} \geq 0$ for all t.

As the dynamics is time-invariant, the policy parameter ϕ reduces to two constant vectors $\phi^{(1)}, \phi^{(2)} \in \mathbb{R}^3_+$. Moreover, because the initial x_0 lies in either Γ_1 or Γ_2, and as explained earlier the state process driven by the parameterized policy (17)–(18) will never escape from the initial region, it suffices to parameterize and update only one of these vectors, which we denote by ϕ.

We run Algorithm 1 to generate iterates of $(\phi, w) \in \mathbb{R}^3_+ \times \mathbb{R}$, and compare the final estimates with the corresponding theoretical values. We emphasize that Algorithm 1 is completely model-free, which uses policy gradient to update ϕ and stochastic approximation to update w.

Results for different expected returns. Figure 1 shows the iterates of ϕ, w, $\mathbb{E}[x_T]$, and $\mathrm{Var}(x_T)$ for $z = 1.10$, 1.15, and 1.20 respectively. Table 1 reports the final estimates. These results demonstrate that our algorithm leads to convergence to the ground truth values without the knowledge of model parameters. In particular, it learns the optimal values of the terminal mean and variance remarkably well.

Extreme target return. We next consider an extreme case $z = 2.0$, corresponding to a 100% expected return within a year. Figure 2 and Table 1 summarize the results. Note that Algorithm 1 enforces the constraint $\phi \geq 0$ using the projected gradient descent method, ensuring that the mean strategy u^ϕ (18) always satisfies the no-shorting constraint. In this case, in terms of the finally learned values, ϕ is significantly away from the ground truth, while the other parameters, especially the terminal mean and variance, are still very close to the respective theoretical values. In other words, the portfolio policy learned from our algorithm is quite different from the theoretically optimal one, but it is nearly optimal because the resulting performance is close to that of the oracle. One must note that in any reinforcement learning problem, what matters is the achieved *performance* much more than the proximity of the learned policy from the optimal one. On the other hand, the high target return renders a very high variance (risk), consistent with the classical mean–variance theory.

Robustness check on initialization and learning rate. The above reports the numerical results with initial values $(\theta, \phi, w) = (0, (0.1, 0.1, 0.1), z)$, and learning rates $(\alpha_\theta, \alpha_\phi, \alpha_w) = (0.001, 0.1, 0.01)$. We have also conducted experiments with different initial points and different learning rates. We do not include the corresponding figures and tables here for a concise numerical section, but just report that the algorithm is robust to initialization. For learning rates, the observed behavior follows the usual pattern of SGD: large learning rates give faster convergence but larger variance, while smaller learning rates lead to slower convergence but smaller variance.

Table 1: Final estimates for different expected returns z, including the extreme case $z = 2.0$.

z	ϕ	ϕ^*	w	w^*	$\mathbb{E}[x_T]$	$\mathbb{E}[x_T^*]$	$\mathrm{Var}(x_T)$	$\mathrm{Var}(x_T^*)$
1.10	(0.090, 1.199, 0.858)	(0.000, 1.341, 0.827)	1.406	1.393	1.101	1.100	0.030	0.029
1.15	(0.015, 1.321, 0.823)	(0.000, 1.341, 0.827)	1.595	1.590	1.150	1.150	0.066	0.066
1.20	(0.017, 1.338, 0.781)	(0.000, 1.341, 0.827)	1.792	1.787	1.197	1.200	0.114	0.117
2.00	(0.189, 1.010, 0.898)	(0.000, 1.341, 0.827)	4.965	4.934	1.957	2.000	2.742	2.934

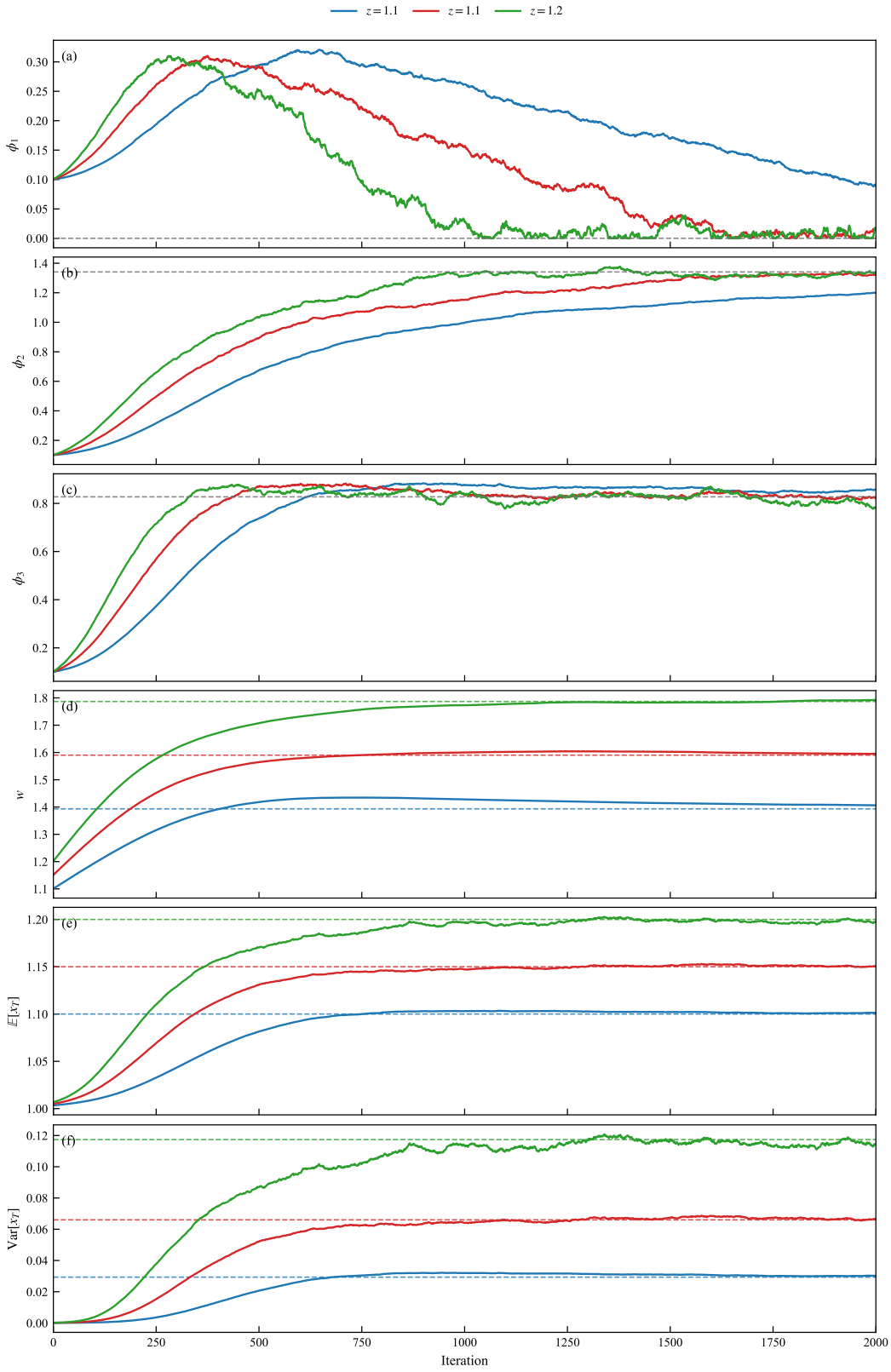

Figure 1: Iterates of ϕ, w, $\mathbb{E}[x_T]$, and $\mathrm{Var}(x_T)$ in the main loop of Algorithm 1 for different expected returns z. For clarity, the first three panels depict the three components of ϕ respectively. Dashed horizontal lines indicate the corresponding theoretical values.

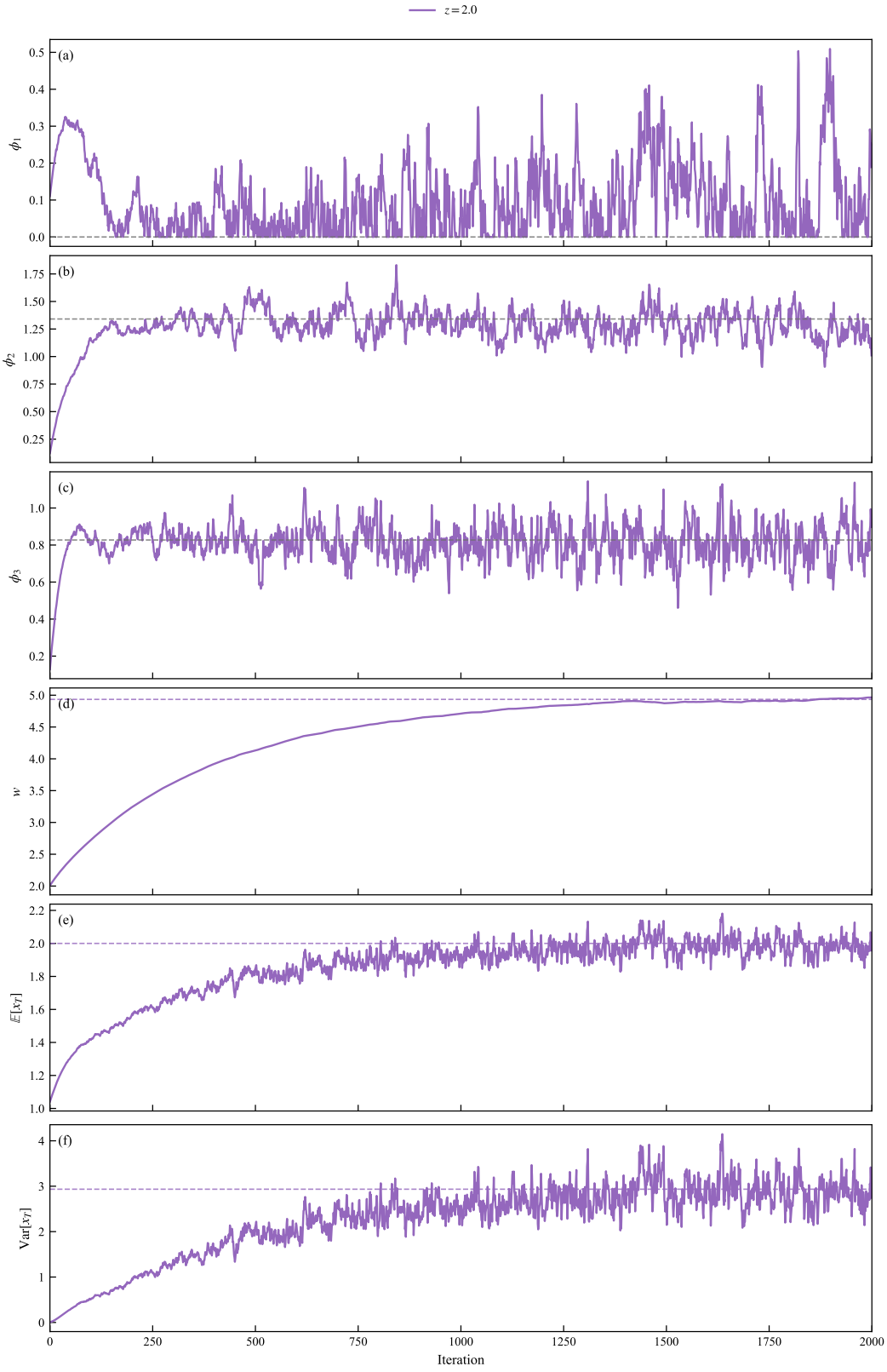

Figure 2: Iterates of ϕ, w, $\mathbb{E}[x_T]$, and $\mathrm{Var}(x_T)$ in the main loop of Algorithm 1 for $z = 2.0$.

6 Conclusions

This paper is the first to treat *constrained* continuous-time MV portfolio selection in the realm of reinforcement learning. The main "trick" is to take an alternative way of randomization: adding suitably chosen Gaussian noises to the admissible controls, instead of randomizing over the constrained control space. The resulting individually sampled policies may violate the constraint, but their mean is automatically feasible which will be improved until ultimately executed and implemented. This approach was first proposed in Dai et al. (2025) and has been applied in different application domains such as generative AI diffusion models (Huang et al., 2026) and image generations (Huang and Zhou, 2026) for *unconstrained* problems, but it is potentially more powerful, as foreshadowed by the present paper, for constrained and other types of RL problems.

References

Beck, A. (2017). *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM.

Bishop, C. M. and Bishop, H. (2024). *Deep Learning: Foundations and Concepts*. Springer.

Boyd, S. and Vandenberghe, L. (2023). *Convex Optimization*. Cambridge University Press, Cambridge.

Chau, H., Nguyen, D., and Nguyen, T. (2026). Continuous-time optimal investment with portfolio constraints: A reinforcement learning approach. *European Journal of Operational Research*, 328(3):1068–1092.

Chen, S., Li, X., and Zhou, X.Y. (1998). Stochastic linear quadratic regulators with indefinite control weight costs. *SIAM Journal on Control and Optimization*, 36:1685–1702.

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

Guo, X., Li, X., and Xu, R. (2026). Fast policy learning for linear-quadratic control with entropy regularization. *SIAM Journal on Control and Optimization*, 64:124–151.

Hu, Y. and Zhou, X. Y. (2005). Constrained stochastic LQ control with random coefficients, and application to portfolio selection. *SIAM Journal on Control and Optimization*, 44:444–466.

Huang, Y., Jia, Y., and Zhou, X. Y. (2024). Mean–variance portfolio selection by continuous-time reinforcement learning: Algorithms, regret analysis, and empirical study. arXiv:2412.16175, 2024.

Huang, Y., Tang, W., and Zhou, X.Y. (2026). ART for diffusion sampling: Continuous-time control and actor–critic Learning. arXiv:2607.02137, 2026.

Haung, Y. and Zhou, X. Y. (2026). Amortized guidance for image inpainting with pretrained diffusion models. arXiv:2605.13010, 2026.

Huang, Y.-J., Wang, Z., and Zhou, Z. (2025). Convergence of policy iteration for entropy-regularized stochastic control problems. *SIAM Journal on Control and Optimization*, 63:752–777.

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

Jia, Y. and Zhou, X. Y. (2022b). Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23:1–50.

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

Li, X. and Xu, Z. Q. (2016). Continuous-time Markowitz's model with constraints on wealth and portfolio. *Operations Research Letters*, 44(6):729–736.

Li, X., Zhou, X. Y., and Lim, A. E. B. (2002). Dynamic mean–variance portfolio selection with no-shorting constraints. *SIAM Journal on Control and Optimization*, 40:1540–1555.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Sethi, D., Šiška, D., and Zhang, Y. (2025). Entropy annealing for policy mirror descent in continuous time and space. *SIAM Journal on Control and Optimization*, 63(4):3006–3041.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, 2nd edition.

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

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

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

Yong, J. and Zhou, X. Y. (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer.

Zhou, X. Y. and Li, D. (2000). Continuous-time mean–variance portfolio selection: A stochastic LQ framework. *Applied Mathematics & Optimization*, 42:19–33.

Appendix

A Viscosity Solutions

For reader's convenience, we recall the standard definition of viscosity solutions used in the proof of Theorem 1; see Yong and Zhou (1999).

When $\Sigma(t,x) = \Psi_t(x-w)^2$, the HJB equation (8) becomes

$$\begin{cases} \dfrac{\partial}{\partial t}v(t,x) + \inf\limits_{a\in\mathbb{R}^m_+} G\left(t,x,a,\dfrac{\partial v}{\partial x},\dfrac{\partial^2 v}{\partial x^2}\right) = 0, \\ v(T,x) = (x-w)^2 - (w-z)^2, \end{cases} \tag{27}$$

where

$$G(t,x,a,p,P) := pa^\intercal\sigma_t\rho_t + \frac{1}{2}P\left[a^\intercal\sigma_t\sigma_t^\intercal a + (x-w)^2\operatorname{tr}(\Psi_t\sigma_t\sigma_t^\intercal)\right].$$

Definition A.1. *A function $v \in C([0,T]\times\mathbb{R})$ is called a* viscosity solution *of (27) if*

$$v(T,x) = (x-w)^2 - (w-z)^2, \qquad \forall x\in\mathbb{R},$$

and for all $(t,x)\in[0,T]\times\mathbb{R}$,

$$\begin{aligned} q + \inf_{a\geq 0} G(t,x,a,p,P) &\geq 0, \qquad && \forall (q,p,P)\in D^{1,2,+}_{t,x}v(t,x), \\ q + \inf_{a\geq 0} G(t,x,a,p,P) &\leq 0, && \forall (q,p,P)\in D^{1,2,-}_{t,x}v(t,x), \end{aligned}$$

where

$$\begin{aligned} D^{1,2,+}_{t,x}v(t_0,x_0) := \Bigg\{\Bigg(&\frac{\partial\varphi}{\partial t}(t_0,x_0),\frac{\partial\varphi}{\partial x}(t_0,x_0),\frac{\partial^2\varphi}{\partial x^2}(t_0,x_0)\;\Bigg|\;\varphi\in C^{1,2}([0,T]\times\mathbb{R}) \text{ and} \\ &v-\varphi \text{ attains a local maximum at } (t_0,x_0)\Bigg)\Bigg\}, \\ D^{1,2,-}_{t,x}v(t_0,x_0) := \Bigg\{\Bigg(&\frac{\partial\varphi}{\partial t}(t_0,x_0),\frac{\partial\varphi}{\partial x}(t_0,x_0),\frac{\partial^2\varphi}{\partial x^2}(t_0,x_0)\;\Bigg|\;\varphi\in C^{1,2}([0,T]\times\mathbb{R}) \text{ and} \\ &v-\varphi \text{ attains a local minimum at } (t_0,x_0)\Bigg)\Bigg\}. \end{aligned}$$

Remark. If $v\in C^{1,2}_{t,x}([0,T]\times\mathbb{R})$, then

$$\begin{aligned} D^{1,2,+}_{t,x}v(t_0,x_0) &= \left\{\frac{\partial v}{\partial t}(t_0,x_0)\right\}\times\left\{\frac{\partial v}{\partial x}(t_0,x_0)\right\}\times\left[\frac{\partial^2 v}{\partial x^2}(t_0,x_0),\infty\right), \\ D^{1,2,-}_{t,x}v(t_0,x_0) &= \left\{\frac{\partial v}{\partial t}(t_0,x_0)\right\}\times\left\{\frac{\partial v}{\partial x}(t_0,x_0)\right\}\times\left(-\infty,\frac{\partial^2 v}{\partial x^2}(t_0,x_0)\right]. \end{aligned}$$

In this case,

$$D^{1,2,+}_{t,x}v(t_0,x_0)\cap D^{1,2,-}_{t,x}v(t_0,x_0) = \left(\frac{\partial v}{\partial t}(t_0,x_0),\frac{\partial v}{\partial x}(t_0,x_0),\frac{\partial^2 v}{\partial x^2}(t_0,x_0)\right).$$