

ON OPTIMAL TRACKING PORTFOLIO IN INCOMPLETE MARKETS: THE REINFORCEMENT LEARNING APPROACH*

LIJUN BO[†], YIJIE HUANG[‡], AND XIANG YU[§]

Abstract. This paper studies an infinite horizon optimal tracking portfolio problem using capital injection in incomplete market models. The benchmark process is modeled by a geometric Brownian motion with zero drift driven by some unhedgeable risk. The relaxed tracking formulation is adopted where the fund account is compensated by the injected capital needs to outperform the benchmark process at any time, and the goal is to minimize the cost of the discounted total capital injection. When model parameters are known, we formulate the equivalent auxiliary control problem with reflected state dynamics, for which the classical solution of the HJB equation with Neumann boundary condition is obtained explicitly. When model parameters are unknown, we introduce the exploratory formulation for the auxiliary control problem with entropy regularization and develop the continuous-time q-learning algorithm in models of reflected diffusion processes. In some illustrative numerical examples, we show the satisfactory performance of the q-learning algorithm.

Key words. optimal tracking portfolio, capital injection, incomplete market, reflected diffusion process, continuous-time q-learning

MSC codes. 93E20, 49J55, 68T05

DOI. 10.1137/23M1620892

1. Introduction. One important business in fund management is to choose the portfolio among some risky assets to closely track some benchmark processes such as the market index, the inflation rates, the exchange rates, the liability, or the living cost and education cost, etc. How to formulate the tracking procedure and derive the optimal tracking portfolio has become an active research topic in quantitative finance during the past decades. For example, several objectives for active portfolio management are proposed in Browne (1999a), Browne (1999b), and Browne (2000) including: maximizing the probability that the agent's wealth achieves a performance goal related to the benchmark before falling below a predetermined shortfall; minimizing the expected time to reach the performance goal; the mixture of these two objectives, and some further extensions by considering the expected reward or expected penalty. Another standard method to optimize the tracking error is to minimize the variance or downside variance relative to the index value or return, which leads to the linear-quadratic stochastic control problem; See Gaivoronski, Krylov, and Wijst (2005), Yao, Zhang, and Zhou (2006), and Ni et al. (2022) among others. In Strub and Baumann (2018), a different objective function is introduced to measure the similarity between

*Received by the editors November 30, 2023; accepted for publication (in revised form) October 30, 2024; published electronically January 23, 2025.

<https://doi.org/10.1137/23M1620892>

Funding: The first and second authors are supported by National Natural Science of Foundation of China (grant 12471451), by the Natural Science Basic Research Program of Shaanxi (grant 2023-JC-JQ-05), and by the Shaanxi Fundamental Science Research Project for Mathematics and Physics (grant 23JSZ010). The third author is supported by the Hong Kong RGC General Research Fund (GRF) under grants 15306523 and 15211524.

[†]School of Mathematics and Statistics, Xidian University, Xi'an, 710126, China (lijunbo@ustc.edu.cn).

[‡]School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui, 230026, China (huang1@mail.ustc.edu.cn).

[§]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (xiang.yu@polyu.edu.hk).

the normalized historical trajectories of the tracking portfolio and the index where the rebalancing transaction costs can also be taken into account. Recently, another new tracking formulation using the fictitious capital injection was studied in Bo, Liao, and Yu (2021), and Bo, Huang, and Yu (2023, 2024), however, all in the complete market model. With the help of an auxiliary control problem based on the state process with reflections, the original stochastic control problem under dynamic floor constraints is equivalent to the study of the HJB equation with a Neumann boundary condition. The market completeness plays a key role therein as the dual transform of the HJB equation leads to a linear PDE problem whose solution admits some probabilistic representations. The existence of the classical solution for the linear dual PDE and the verification of the feedback optimal portfolio control can be obtained by using stochastic flow analysis and some delicate estimations.

The present paper aims to further extend the studies on the relaxed tracking portfolio using capital injection in Bo, Liao, and Yu (2021) and Bo, Huang, and Yu (2023, 2024) in some incomplete market models in which there exists some unhedgeable risk driving the external benchmark process (see its definition in (2.4)). As a direct consequence, the previous methodology based on the dual transform and the probabilistic representations for the dual PDE therein fails because the dual HJB equation can no longer be linearized. We consider the benchmark process governed by a geometric Brownian motion with zero drift in the present paper. With knowledge of market parameters, we formulate an equivalent auxiliary stochastic control problem similar to Bo, Liao, and Yu (2021), which results in a stochastic control problem with the underlying state process exhibiting reflections at the boundary zero. Leveraging the benchmark dynamics as a geometric Brownian motion facilitates the dimension reduction of the control problem. We then derive the explicit classical solution to the associated HJB equation with a Neumann boundary condition. Furthermore, we consider the realistic situation when all market model parameters are unknown, for which we are interested in developing the continuous-time reinforcement learning approach. Recently, for continuous-time stochastic control problems, Wang, Zariphopoulou, and Zhou (2020), Tang Zhang and Zhou (2022), Jia and Zhou (2022a, 2022b, 2023) have developed the theoretical foundation for reinforcement learning in the continuous-time exploration formulation with continuous state space and action space. In particular, Wang, Zariphopoulou, and Zhou (2020) first studied the exploratory learning formulation by incorporating the entropy regularization to encourage the policy exploration, and the optimal policy has been shown as a Gaussian type for Linear-Quadratic (LQ) control problems. Jia and Zhou (2022a) examined the policy evaluation problem and proposed the martingale loss to facilitate the algorithm design. The policy gradient problem was then studied in Jia and Zhou (2022b) where the martingale approach in Jia and Zhou (2022a) can be adopted for the policy gradient and some actor-critic algorithms can be devised to learn the value function and the stochastic policies alternatively. Later, Jia and Zhou (2023) established a continuous time q-learning theory by considering the first order approximation of the conventional discrete-time Q-function. We also refer to some recent subsequent studies in different contexts such as Wang, Gao, and Li (2023) in applying the continuous-time actor-critic algorithm for solving the optimal execution problem in the Almgren–Chriss model; the application of a policy improvement algorithm in Bai et al. (2023) for solving the optimal dividend problem in diffusion models; the generalization of continuous-time q-learning in Wei and Yu (2025) for mean-field control problems in McKean–Vlasov diffusion models. The occupation measure considered in Zhao, Tang, and Yao (2023)

facilitates the derivation of performance-difference and local-approximation formulas for continuous-time policy gradient and trust region policy optimization algorithms.

In this paper, we are interested in generalizing the continuous-time q-learning from the standard diffusion models in Jia and Zhou (2023) to solve our auxiliary stochastic control problem with the reflection boundary at 0. In particular, for the stochastic control problems with reflections, it is shown in the present paper that the value function and the associated q-function can also be characterized by the martingale condition of some stochastic processes involving the reflection term or the local time at 0; see Proposition 4.2 and Theorem 4.3. In addition, some extra transversality conditions (see (4.16) and (4.10)) are also needed in our infinite horizon setting. Building upon our martingale characterization in Theorem 4.3, we are able to devise the offline q-learning algorithm for the targeted stochastic control problems with reflections using the stochastic approximation resulting from the martingale orthogonality condition. As a new theoretical contribution to the literature, in the context of the infinite horizon stochastic control problems with reflections, we establish the convergence of the stochastic approximation algorithms for time discretization and horizon truncation when $\Delta t \rightarrow 0$ and $T \rightarrow \infty$; see Theorem 4.5. To illustrate the efficiency of our q-learning algorithm, we study a simulation example of our optimal tracking portfolio problem with a particular choice of the temperature parameter $\gamma = \rho/d$. Here ρ stands for the subjective discount rate and d is the dimension of stocks, both are known to the agent. In this case, we present an example beyond the LQ control framework such that the exploratory HJB equation with a Neumann boundary condition can be solved fully explicitly, allowing us to obtain the exact parameterization of the optimal value function and the optimal q-function. For some initial inputs and choices of learning rates, we illustrate the very satisfactory performance of the iteration convergence of learned parameters towards their true values.

The remainder of this paper is organized as follows. Section 2 introduces the incomplete market model and the benchmark process as well as the portfolio optimization problem under the relaxed tracking formulation. Section 3 delves into the classical control problem when the model is known. By adopting an auxiliary control formulation and conducting dimension reduction, the HJB equation is solved explicitly. In section 4, when the market parameters are unknown, the continuous-time reinforcement learning approach is proposed for the exploratory formulation of the control problem with entropy regularization. In particular, the q-learning algorithm is developed therein together with some convergence results on time discretization and horizon truncation. Section 5 studies some numerical examples by implementing the q-learning algorithm. In addition, we present the comparison between the q-learning algorithm and the maximum likelihood estimation (MLE) for some real market data to demonstrate the outperformance of our algorithm. Finally, all proofs of the main results in previous sections are collected in section 6.

2. Market model and problem formulation. Let $(\Omega, \mathcal{F}, \mathbb{F}^W, \mathbb{P}^W)$ be a filtered probability space with the filtration $\mathbb{F}^W = (\mathcal{F}_t^W)_{t \geq 0}$ satisfying the usual conditions, which supports a $(d+1)$ -dimensional Brownian motion $(W^0, W^1, \dots, W^d) = (W_t^0, W_t^1, \dots, W_t^d)_{t \geq 0}$. We consider a financial market consisting of d risky assets whose price dynamics are driven by the d -dimensional Brownian motion (W^1, \dots, W^d) satisfying

$$(2.1) \quad \frac{dS_t^i}{S_t^i} = \mu_i dt + \sum_{j=1}^d \sigma_{ij} dW_t^j, \quad i = 1, \dots, d, \quad t \geq 0,$$

with the mean return rate $\mu_i \in \mathbb{R}$ and the volatility $\sigma_{ij} \in \mathbb{R}$. Hereafter, we denote by $\mu = (\mu_1, \dots, \mu_d)^\top$ the vector of return rate and by $\sigma = (\sigma_{ij})_{d \times d}$ the volatility matrix. Assume that the riskless interest rate $r = 0$ that amounts to the change of numéraire and the return rate $\mu \neq \mathbf{0}$. From this point onwards, all processes including the wealth process and the benchmark process are defined after the change of numéraire.

For $t \geq 0$, let θ_t^i be the amount of wealth (the resulting process $\theta^i = (\theta_t^i)_{t \geq 0}$ is assumed to be \mathbb{F}^W -adapted) that the fund manager allocates in asset $S^i = (S_t^i)_{t \geq 0}$ at time t . The self-financing wealth process under the portfolio control $\theta = (\theta_t^1, \dots, \theta_t^d)_{t \geq 0}^\top$ is given by, for $t \geq 0$,

$$(2.2) \quad V_t^\theta = v + \int_0^t \theta_s^\top \mu ds + \int_0^t \theta_s^\top \sigma dW_s, \quad V_0^\theta = v \in \mathbb{R}_+ := [0, \infty).$$

We consider the portfolio decision making by the fund manager whose goal is to optimally track a stochastic benchmark process $Z = (Z_t)_{t \geq 0}$. In the present paper, we consider the benchmark process satisfying the geometric Brownian motion with zero drift as studied in Björk and Landén (2000) that may refer to the futures price process of a commodity, a stock index, or an interest rate index. As in Björk and Landén (2000), the benchmark process $Z = (Z_t)_{t \geq 0}$ is assumed to satisfy the form of

$$(2.3) \quad \frac{dZ_t}{Z_t} = \sigma_Z dW_t^\kappa \quad \forall t > 0$$

with the initial value $Z_0 = z > 0$ and the volatility $\sigma_Z > 0$.

For the nonzero correlative coefficient $\kappa \in [-1, 1]$, the process $W^\kappa = (W_t^\kappa)_{t \geq 0}$ appearing in (2.3) is a standard Brownian motion, which is defined by

$$(2.4) \quad W_t^\kappa := \kappa W_t^0 + \sqrt{1 - \kappa^2} W_t^\eta, \quad t \geq 0.$$

Here $W^\eta = (W_t^\eta)_{t \geq 0}$ is a linear combination of d -dimensional Brownian motion $W = (W^1, \dots, W^d)$ with weights $\eta = (\eta_1, \dots, \eta_d)^\top \in [-1, 1]^d$. We also recall that $W^0 = (W_t^0)_{t \geq 0}$ is a Brownian motion independent of the d -dimensional Brownian motion W , which stands for the unhedgeable risk. As a consequence, the market model is incomplete.

Given the benchmark process $Z = (Z_t)_{t \geq 0}$, an optimal tracking portfolio problem is considered that combines the portfolio control with another singular control as capital injection together with dynamic floor constraints. To be precise, we assume that the fund manager can strategically inject capital $A = (A_t)_{t \geq 0}$ into the fund account from time to time whenever it is necessary such that the total capital dynamically dominates the benchmark process Z , i.e., $A_t + V_t^\theta \geq Z_t$ for all $t \geq 0$. The goal of the optimal tracking problem is to minimize the expected cost of the discounted total capital injection under dynamic floor constraints that

$$(2.5) \quad w(v, z) := \inf_{A, \theta} \mathbb{E} \left[A_0 + \int_0^\infty e^{-\rho t} dA_t \right] \quad \text{s.t.} \quad Z_t \leq A_t + V_t^\theta \text{ at each } t \geq 0,$$

where the constant $\rho > 0$ is the discount rate and $A_0 = (z - v)^+$ is the initial injected capital to match with the initial benchmark.

To cope with the problem (2.5) with dynamic floor constraints, we observe that, for a fixed control θ , the optimal A is the smallest adapted right-continuous and nondecreasing process that dominates $Z - V^\theta$. Let \mathbb{U} be the set of regular \mathbb{F} -adapted control processes $\theta = (\theta_t)_{t \geq 0}$ such that (2.2) is well-defined. Similarly to Bo, Liao, and

Yu (2021), for each fixed regular control θ , the optimal singular control A_t^* satisfies that $A_t^* = 0 \vee \sup_{s \in [0, t]} (Z_s - V_s^\theta) \forall t \geq 0$. As a result, the problem (2.5) with floor constraints admits the equivalent formulation as an unconstrained control problem but with a running maximum cost that

$$(2.6) \quad w(v, z) = (z - v)^+ + \inf_{\theta \in \mathbb{U}} \mathbb{E} \left[\int_0^\infty e^{-\rho t} d \left(0 \vee \sup_{s \in [0, t]} (Z_s - V_s^\theta) \right) \right].$$

In the next section, we will focus on the solvability of problem (2.6) in the sense of strong control by introducing an auxiliary control problem.

3. The auxiliary control problem. To formulate the auxiliary control problem, we impose a new controlled state process to replace the process $V^\theta = (V_t^\theta)_{t \geq 0}$ given in (2.2). To do it, we first define the following difference process by $D_t := Z_t - V_t^\theta + v - z \forall t \geq 0$. It is obvious that $D_0 = 0$. Moreover, for any $x \geq 0$, let us consider the running maximum process of $D = (D_t)_{t \geq 0}$ defined by

$$(3.1) \quad L_t := x \vee \sup_{s \in [0, t]} D_s \geq 0 \quad \forall t > 0$$

with the initial value $L_0 = x \geq 0$. One can easily see that $(z - v)^+ - w(z, v)$ with the value function $w(v, z)$ given in (2.6) is equivalent to the auxiliary control problem,

$$(3.2) \quad \sup_{\theta \in \mathbb{U}} \mathbb{E} \left[- \int_0^\infty e^{-\rho t} dL_t \right],$$

when we set the initial level $L_0 = x = (v - z)^+$.

We start with the introduction of a new controlled state process $X = (X_t)_{t \geq 0}$ for problem (3.2), which is defined as the reflected process $X_t := L_t - D_t$ for $t \geq 0$ that satisfies the SDE

$$(3.3) \quad X_t = - \int_0^t \sigma_Z Z_s dW_s^\kappa + \int_0^t \theta_s^\top \mu ds + \int_0^t \theta_s^\top \sigma dW_s + L_t \quad \forall t \geq 0$$

with the initial value $X_0 = x \geq 0$. In particular, the running maximum process L_t increases if and only if $X_t = 0$, i.e., $L_t = D_t$. We will change the notation from L_t to L_t^X from this point onwards to emphasize its dependence on the new state process X given in (3.3). The benchmark process $Z = (Z_t)_{t \geq 0}$ defined in (2.3) is chosen as another state process.

Let \mathbb{U}^r be the set of admissible portfolios (controls) such that the reflected SDE (3.3) has a unique strong solution. Then, we propose the following stochastic control problem, for $(x, z) \in \mathbb{R}_+ \times (0, \infty)$,

$$(3.4) \quad \hat{w}(x, z) := \sup_{\theta \in \mathbb{U}^r} J(\theta; x, z) := \sup_{\theta \in \mathbb{U}^r} \mathbb{E}_{x, z} \left[- \int_0^\infty e^{-\rho t} dL_t^X \right],$$

where $\mathbb{E}_{x, z}[\cdot] := \mathbb{E}[\cdot | X_0 = x, Z_0 = z]$. Using the definition (3.4) of $\hat{w}(x, z)$, it is not difficult to check that (i) the value function $x \rightarrow \hat{w}(x, z)$ is nondecreasing; (ii) it holds that $\sup_{x_1, x_2 \geq 0} \left| \frac{\hat{w}(x_1, z) - \hat{w}(x_2, z)}{x_1 - x_2} \right| \leq 1$, that is, the value function $x \rightarrow \hat{w}(x, z)$ is Lipschitz continuous with Lipschitz coefficient being 1. These properties will be applied in the remaining sections.

We next apply the change of measure to simplify the stochastic control problem in (3.4) by reducing the dimension. To this end, let us introduce the normalized processes of triplet (X, θ, L^X) by the benchmark process $Z = (Z_t)_{t \geq 0}$ defined by

$$(3.5) \quad Y_t := \frac{X_t}{Z_t}, \quad \tilde{\theta}_t := \frac{\theta_t}{Z_t}, \quad L_t^Y := \int_0^t \frac{dL_s^X}{Z_s} \quad \forall t \geq 0.$$

Note that the process Z is strictly positive (see (2.3)). Then, $t \rightarrow L_t^Y$ is a nondecreasing process and satisfies that, a.s., $\int_0^t \mathbf{1}_{Y_s=0} dL_s^Y = \int_0^t \mathbf{1}_{X_s=0} \frac{dL_s^X}{Z_s} = \int_0^t \frac{dL_s^X}{Z_s} = L_t^Y$ for $t \geq 0$. This implies that $L^Y = (L_t^Y)_{t \geq 0}$ is the local time process for the process $Y = (Y_t)_{t \geq 0}$ at the reflecting boundary 0. Then, Itô's rule yields that

$$(3.6) \quad dY_t = \sigma_Z^2(Y_t + 1)dt - \sigma_Z(Y_t + 1)dW_t^\kappa + \tilde{\theta}_t^\top \mu dt + \tilde{\theta}_t^\top \sigma dW_t + dL_t^Y.$$

In the following, we introduce the following change of measure specified by

$$(3.7) \quad \frac{d\mathbb{Q}^W}{d\mathbb{P}^W} \Big|_{\mathcal{F}_t} = Z_t \quad \forall t \geq 0.$$

We then consider the following stochastic control problem formulated as

$$(3.8) \quad u(y) := \sup_{\theta \in \tilde{\mathbb{U}}^r} \mathbb{E}^{\mathbb{Q}^W} \left[- \int_0^\infty e^{-\rho s} dL_s^Y \right] \quad \forall y \in \mathbb{R}_+,$$

where the normalized admissible set $\tilde{\mathbb{U}}^r := \{(\frac{\theta_t}{Z_t})_{t \geq 0}; \theta \in \mathbb{U}^r\}$, and $\mathbb{E}^{\mathbb{Q}^W}[\cdot]$ is the expectation under the probability measure $\mathbb{Q}^W \sim \mathbb{P}^W$. It follows from (3.6) that the underlying state process Y has the following dynamics under \mathbb{Q}^W given by

$$(3.9) \quad dY_t = -\sigma_Z(Y_t + 1)d\tilde{W}_t^\kappa + \tilde{\theta}_t^\top \tilde{\mu} dt + \tilde{\theta}_t^\top \sigma d\tilde{W}_t + dL_t^Y.$$

Here $\tilde{\mu} := \mu + \sqrt{1 - \kappa^2} \sigma_Z \sigma \eta$, $\tilde{W}_t^\kappa := W_t^\kappa - \sigma_Z t$ and $\tilde{W}_t := W_t - \sqrt{1 - \kappa^2} \sigma_Z t \eta$ for $t \geq 0$ are \mathbb{Q}^W -Brownian motions. For the case $\kappa = 1$, the process $\tilde{W}^\kappa = (\tilde{W}_t^\kappa)_{t \geq 0}$ is also independent of $\tilde{W} = W = (W^1, \dots, W^d)^\top$. The following lemma gives the relationship between the value function $u(y)$ for $y \geq 0$ defined by (3.8) and the value function $\hat{w}(x, z)$ for $(x, z) \in \mathbb{R}_+ \times (0, \infty)$ defined by (3.4), whose proof is straightforward and hence omitted.

LEMMA 3.1. *Let $(x, z) \in \mathbb{R}_+ \times (0, \infty)$. It holds that $\hat{w}(x, z) = zu(\frac{x}{z})$.*

Lemma 3.1 tells us that we can solve the simplified stochastic control problem (3.8) with a one-dimensional state process with reflection instead of problem (3.4). Then, the corresponding value function satisfies the following HJB equation given by, for $y > 0$,

$$(3.10) \quad \sup_{\tilde{\theta} \in \mathbb{R}^n} \left\{ u'(y) \tilde{\theta}^\top \tilde{\mu} + u''(y) \tilde{\theta}^\top \sigma \sigma^\top \tilde{\theta} - \sqrt{1 - \kappa^2} \sigma_Z (y + 1) \tilde{\theta}^\top \sigma \eta u''(y) \right\} + \frac{1}{2} \sigma_Z^2 (y + 1)^2 u''(y) = \rho u(y)$$

with Neumann boundary condition $u'(0) = 1$. If the value function u is strictly concave, then the optimal feedback portfolio is given by

$$(3.11) \quad \tilde{\theta}^*(y) = -(\sigma \sigma^\top)^{-1} \frac{u'(y) \tilde{\mu} - \sqrt{1 - \kappa^2} \sigma_Z (y + 1) \sigma \eta u''(y)}{u''(y)} \quad \forall y \geq 0.$$

By plugging (3.11) into (3.10), we have

$$(3.12) \quad -\alpha \frac{(u'(y))^2}{u''(y)} + \sqrt{1 - \kappa^2} \zeta (y + 1) u'(y) + \frac{1}{2} \sigma_Z^2 \kappa^2 (y + 1)^2 u''(y) = \rho u(y),$$

where the coefficients are defined by $\alpha := \frac{1}{2} \tilde{\mu}^\top (\sigma \sigma^\top)^{-1} \tilde{\mu}$ and $\zeta := \sigma_Z \eta^\top \sigma^{-1} \tilde{\mu}$. We consider the candidate solution of (3.12) given by

$$(3.13) \quad u(y) = \frac{\lambda - 1}{\lambda} (1 + y)^{\frac{\lambda}{\lambda - 1}} \quad \forall y \geq 0.$$

Here, $\lambda \in (0, 1)$ is the unique solution to the following equation:

$$(3.14) \quad \ell(\lambda) := \alpha\lambda(\lambda-1)^2 + \rho(\lambda-1)^2 - \sqrt{1-\kappa^2}\zeta\lambda(\lambda-1) - \frac{1}{2}\kappa^2\sigma_Z^2\lambda = 0.$$

Hence, (3.14) has a unique root $\lambda \in (0, 1)$. Based on the above results, we have the following verification result for the strong control problem, whose proof is standard and hence omitted.

PROPOSITION 3.2. *The function $u(y)$ for $y \geq 0$ given by (3.13) is a classical solution of the HJB equation (3.10). Define the following optimal feedback control function by, for all $y \in \mathbb{R}_+$,*

$$(3.15) \quad \tilde{\theta}^*(y) = (1-\lambda)(y+1)(\sigma\sigma^\top)^{-1}\tilde{\mu} + \sqrt{1-\kappa^2}\sigma_Z(y+1)(\sigma\sigma^\top)^{-1}\sigma\eta.$$

Consider the controlled state process $Y^* = (Y_t^*)_{t \geq 0}$ that obeys the following SDE, for all $t \geq 0$,

$$(3.16) \quad Y_t^* = - \int_0^t \sigma_Z(Y_s + 1) d\tilde{W}_s^\kappa + \int_0^t \tilde{\theta}^*(Y_s^*)^\top \tilde{\mu} ds + \int_0^t \tilde{\theta}^*(Y_s^*)^\top \sigma d\tilde{W}_s + L_t^{Y^*}$$

with $L_0^{Y^*} = y$. Let $\tilde{\theta}_t^* = \tilde{\theta}^*(Y_t^*)$ for all $t \geq 0$. Then $\tilde{\theta}^* = (\tilde{\theta}_t^*)_{t \geq 0} \in \tilde{\mathcal{U}}^r$ is an optimal investment strategy.

4. The continuous-time reinforcement learning approach. In this section, the model is assumed to be unknown to the fund manager, i.e., all model parameters $\mu = (\mu_1, \dots, \mu_d)^\top$, $\sigma = (\sigma_{ij})_{d \times d}$, σ_Z , κ , and $\eta = (\eta_1, \dots, \eta_d)^\top \in [-1, 1]^d$ are unknown. Thereby, the classical control approach in the previous section is not applicable. Our goal is to develop a continuous-time reinforcement learning algorithm to find the optimal tracking portfolio in problem (3.8).

Reinforcement learning allows the decision maker to learn the optimal action in the unknown environment through the interactions with the environment as the repeated trial-and-error procedure. Specifically, the agent exercises a sequel of actions $\tilde{\theta} = (\tilde{\theta}_t)_{t \geq 0}$ and observes responses from $(Y^\theta, L^Y) = (Y_t^\theta, L_t^Y)_{t \geq 0}$ along with a stream of discounted running rewards $(-e^{-\rho t} dL_t^Y)_{t \geq 0}$, and continuously updates and improves his actions based on these observations.

To describe the exploration step in reinforcement learning, we randomize the action $\tilde{\theta}$ and consider its distribution. Assume that the probability space is rich enough to support the uniformly distributed random variable on $[0, 1]$ that is independent of $(\tilde{W}^\kappa, W^1, \dots, W^d)$ and can be used to generate other random variables with specified density functions. Let $K = (K_t)_{t \in [0, T]}$ be a process of mutually independent copies of a uniform random variable on $[0, 1]$ which is also independent of the Brownian motions (BMs) (W^0, W^1, \dots, W^d) , the construction of which requires a suitable extension of probability space (cf. Sun 2006). We then further expand the filtered probability space to $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$, where $\mathbb{F} = (\mathcal{F}_t^W \vee \sigma(K_s, 0 \leq s \leq t))_{t \geq 0}$ and the probability measure \mathbb{Q} , now defined on \mathbb{F} , is an extension from \mathbb{Q}^W (i.e., the two probability measures coincide when restricted to \mathbb{F}^W). In particular, let $\pi: y \in \mathbb{R}_+ \rightarrow \pi(\cdot | y) \in \mathcal{P}(\mathcal{A})$ be a given (feedback) policy with $\mathcal{A} := \mathbb{R}^d$, and $\mathcal{P}(\mathcal{A})$ is a suitable collection of probability density functions. At each time $t \geq 0$, an action $\tilde{\theta}_t^\pi$ is generated from the joint density $\pi(\cdot | Y_t)$. Fixing a policy π and an initial state $y \geq 0$, we can consider the following reflected SDE:

$$(4.1) \quad dY_t^\pi = -\sigma_Z(Y_t^\pi + 1) d\tilde{W}_t^\kappa + (\tilde{\theta}_t^\pi)^\top \tilde{\mu} dt + (\tilde{\theta}_t^\pi)^\top \sigma d\tilde{W}_t + dL_t^{Y^\pi}, \quad t > 0, \quad Y_0^\pi = y.$$

Here, $L^{Y^\pi} = (L_t^{Y^\pi})_{t \geq 0}$ denotes the local time of the state process $Y^\pi = (Y_t^\pi)_{t \geq 0}$ at the level 0.

To encourage exploration, we adopt the Shannon's entropy regularizer suggested in Wang et al. (2020) that leads to the following value function associated to a given policy π that

$$(4.2) \quad J(y; \pi) = \mathbb{E}^{\mathbb{Q}} \left[-\gamma \int_0^\infty e^{-\rho t} \ln \pi(\tilde{\theta}_t^\pi | Y_t^\pi) dt - \int_0^\infty e^{-\rho t} dL_t^{Y^\pi} \middle| Y_0^\pi = y \right],$$

where $\mathbb{E}^{\mathbb{Q}}$ is the expectation with respect to both the Brownian motion and the action randomization. In the above, $\gamma > 0$ is a given parameter indicating the level of exploration, also known as the temperature parameter.

Similarly to Wang, Zariphopoulou, and Zhou (2020), we can show that the average of the sample trajectories $(Y_t^\pi)_{t \geq 0}$ converges to $(\tilde{Y}_t^\pi)_{t \geq 0}$, which satisfies the following reflected SDE,

$$(4.3) \quad d\tilde{Y}_t^\pi = \tilde{b}(\tilde{Y}_t^\pi, \pi(\cdot | \tilde{Y}_t^\pi))dt + \tilde{\sigma}(\tilde{Y}_t^\pi, \pi(\cdot | \tilde{Y}_t^\pi))dB_t + dL_t^\pi,$$

where $B = (B_t)_{t \geq 0}$ is a standard BM independent of (\tilde{W}^κ, W) , the coefficient functions $(\tilde{b}, \tilde{\sigma})$ are, respectively, defined by, for $(y, \pi) \in \mathbb{R}_+ \times \mathcal{P}(\mathcal{A})$,

$$\begin{cases} \tilde{b}(y, \pi) := \int_{\mathcal{A}} \tilde{\theta}^\top \tilde{\mu} \pi(\tilde{\theta} | y) d\tilde{\theta}, \\ \tilde{\sigma}(y, \pi) := \sqrt{\int_{\mathcal{A}} \left(\frac{1}{2} \tilde{\theta}^\top \sigma \sigma^\top \tilde{\theta} - \sqrt{1 - \kappa^2} \sigma_Z (y+1) \tilde{\theta}^\top \sigma \eta + \frac{\sigma_Z^2}{2} (y+1)^2 \right) \pi(\tilde{\theta} | y) d\tilde{\theta}}. \end{cases}$$

Here, $L^\pi = (L_t^\pi)_{t \geq 0}$ denotes the local time process of the state process $\tilde{Y}^\pi = (\tilde{Y}_t^\pi)_{t \geq 0}$ at the level 0, namely, $L^\pi = (L_t^\pi)_{t \geq 0}$ is a continuous nondecreasing process satisfying $\int_0^t \mathbf{1}_{\tilde{Y}_s^\pi = 0} dL_s^\pi = L_t^\pi$ for all $t \geq 0$. It follows from the property of Markovian projection in Brunick and Shreve (2013) that Y_t^π and \tilde{Y}_t^π have the same distribution for any $t \geq 0$. As a consequence, the value function (4.2) is equivalent to the following relaxed control form (see also Kushner and Dupuis (1991) and Kushner (1998)):

$$(4.4) \quad J(y; \pi) = \mathbb{E}^{\mathbb{Q}^W} \left[-\gamma \int_0^\infty e^{-\rho t} \int_{\mathcal{A}} \ln \pi(\tilde{\theta} | \tilde{Y}_t^\pi) \pi(\tilde{\theta} | \tilde{Y}_t^\pi) d\tilde{\theta} dt - \int_0^\infty e^{-\rho t} dL_t^\pi \middle| \tilde{Y}_0^\pi = y \right].$$

The task of reinforcement learning is to find the optimal policy to attain the maximum of the value function

$$(4.5) \quad v(y) = \max_{\pi \in \Pi} J(y; \pi), \quad y \geq 0,$$

where the set Π stands for the set of admissible (stochastic) policies. The following gives the precise definition of admissible policies.

DEFINITION 4.1. A policy π is called admissible, that is, $\pi \in \Pi$, if

- (i) π takes the feedback form as $\pi_t = \pi(\cdot | Y_t)$ for $t \geq 0$, where $\pi(\cdot | \cdot) : \mathcal{A} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ is a measurable function and $\pi(\cdot | y) \in \mathcal{P}(\mathcal{A})$ for all $y \in \mathbb{R}_+$;
- (ii) the SDE (4.3) admits a unique strong solution for any initial $y \in \mathbb{R}_+$.

Assuming that the value function (4.4) under a given admissible policy $\pi \in \Pi$ is smooth enough, we can see that it satisfies the following Neumann problem, for $\pi \in \mathcal{P}(\mathcal{A})$,

$$(4.6) \quad \begin{cases} \int_{\mathcal{A}} \left[H(\tilde{\theta}, y, J_y(y; \pi), J_{yy}(y; \pi)) - \gamma \ln \pi(\tilde{\theta} | y) \right] \pi(\tilde{\theta} | y) d\tilde{\theta} \\ \quad + \frac{\sigma_Z^2}{2} (y+1)^2 J_{yy}(y; \pi) = \rho J(y; \pi), \\ J_y(0; \pi) = 1. \end{cases}$$

On the other hand, the optimal value function defined by (4.5) satisfies the following exploratory HJB equation with a Neumann boundary condition that, for $y \geq 0$,

$$(4.7) \quad \begin{cases} \sup_{\pi \in \mathcal{P}(\mathcal{A})} \int_{\mathcal{A}} \left[H(\tilde{\theta}, y, v_y(y), v_{yy}(y)) - \gamma \ln \pi(\tilde{\theta}) \right] \pi(\tilde{\theta}) d\tilde{\theta} + \frac{\sigma_Z^2}{2} (y+1)^2 v_{yy}(y) = \rho v(y), \\ v_y(0) = 1. \end{cases}$$

Here, the Hamilton operator $H : \mathcal{A} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is defined by

$$(4.8) \quad H(\tilde{\theta}, y, P, Q) := \tilde{\theta}^\top \tilde{\mu} P + \frac{1}{2} \tilde{\theta}^\top \sigma \sigma^\top \tilde{\theta} Q - \sqrt{1 - \kappa^2 \sigma_Z^2} (y+1) \tilde{\theta}^\top \sigma \eta Q.$$

Moreover, by assuming $v_{yy} < 0$, the candidate optimal policy is a Gaussian measure after normalization so that

$$(4.9) \quad \pi^*(\tilde{\theta} | y) = \frac{\exp \left\{ \frac{1}{\gamma} H(\tilde{\theta}, y, v_y(y), v_{yy}(y)) \right\}}{\int_{\mathcal{A}} \exp \left\{ \frac{1}{\gamma} H(\tilde{\theta}, y, v_y(y), v_{yy}(y)) \right\} d\tilde{\theta}}.$$

In what follows, we first establish the policy improvement theorem.

THEOREM 4.1 (policy improvement theorem). *For any given $\pi \in \Pi$, assume that the value function $J(\cdot; \pi) \in C^2(\mathbb{R}_+)$ satisfies (4.6) with the Neumann boundary condition and $J_{yy}(\cdot; \pi) < 0$ for any $y \in \mathbb{R}_+$. We consider the mapping \mathcal{I} on Π defined by*

$$\begin{aligned} \mathcal{I}(\pi) &:= \frac{\exp \left\{ \frac{1}{\gamma} H(\cdot, y, J_y(y; \pi), J_{yy}(y; \pi)) \right\}}{\int_{\mathcal{A}} \exp \left\{ \frac{1}{\gamma} H(\tilde{\theta}, y, J_y(y; \pi), J_{yy}(y; \pi)) \right\} d\tilde{\theta}} \\ &= \mathcal{N} \left(\tilde{\theta} \mid -(\sigma \sigma^\top)^{-1} \frac{\tilde{\mu} J_y(y; \pi) - \sqrt{1 - \kappa^2 \sigma_Z^2} (y+1) \sigma \eta J_{yy}(y; \pi)}{J_{yy}(y; \pi)}, \right. \\ &\quad \left. -(\sigma \sigma^\top)^{-1} \frac{\gamma}{J_{yy}(y; \pi)} \right), \end{aligned}$$

where we denote by $\mathcal{N}(\tilde{\theta} | a, b)$ the Gaussian density function with mean vector $a \in \mathbb{R}^d$ and covariance matrix $b \in \mathbb{R}^{d \times d}$. Then, we have

(i) denoting by $\pi' = \mathcal{I}(\pi)$, if $\pi' \in \Pi$ satisfies

$$(4.10) \quad \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} J(\tilde{Y}_T^{\pi'}; \pi) \right] = 0,$$

then it holds that $J(y; \pi') \geq J(y; \pi)$ for all $y \in \mathbb{R}_+$;

(ii) if the map \mathcal{I} has a fixed point $\pi^* \in \Pi$ satisfying

$$(4.11) \quad \begin{cases} \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} J(\tilde{Y}_T^{\pi^*}; \pi^*) \right] = 0, \\ \limsup_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} J(\tilde{Y}_T^{\pi}; \pi^*) \right] \geq 0 \quad \forall \pi \in \Pi, \end{cases}$$

then π^* is the optimal policy that $v(y) = \max_{\pi \in \Pi} J(y; \pi) = J(y; \pi^*)$;

- (iii) in particular, if we choose the temperature parameter $\gamma = \frac{\rho}{d}$, the map \mathcal{I} has an explicit fixed point $\pi^* \in \Pi$ given by

$$(4.12) \quad \pi^*(\cdot|y) = \mathcal{N}\left(y \mid (\sigma\sigma^\top)^{-1}(\tilde{\mu} + \sigma_z \sqrt{1 - \kappa^2 \sigma \eta})(1+y), (\sigma\sigma^\top)^{-1}\gamma(1+y)^2\right).$$

Theorem 4.1(i) provides a theoretical result for the policy improvement iteration; while Theorem 4.1(ii) shows that the optimal policy is characterized by a fixed point of the iteration operator. In fact, Theorem 4.1(i) holds true for general Hamilton operator H under some mild assumptions. However, it is usually difficult to verify the existence of fixed points and the transversality conditions (4.11) for the general case. To address this issue, we choose a specific temperature parameter $\gamma = \rho/d$ in Theorem 4.1(iii), which significantly simplifies the finding of a fixed point of the mapping \mathcal{I} by using the explicit expression (4.12). The proof of Theorem 4.1(iii) also shows that if we start with a special Gaussian policy as $\pi_0(\cdot|y) \sim \mathcal{N}(\cdot \mid c_1(1+y), c_2(1+y)^2)$, then just after two steps of iterations, the value function can no longer be improved, i.e., the fixed point is attained. We will mainly focus on this special choice of $\gamma = \rho/d$ and verify the transversality conditions (4.11) in section 5, which implies that the fixed point (4.12) is the optimal policy.

Note that the policy improvement iteration in Theorem 4.1 depends on the knowledge of the model parameters, which are not known in the reinforcement learning procedure. Thus, in order to design an implementable algorithm, we turn to generalizing the q-learning theory initially proposed in Jia and Zhou (2023) for our purpose.

4.1. q-function and martingale characterizations. The aim of this section is to derive the q-function of our optimal tracking problem in continuous time and provide martingale characterizations of the q-function. We first give the definition of the q-function as the counterpart of the Q-function in the continuous time framework (cf. Jia and Zhou, 2023).

DEFINITION 4.2 (q-function). *The q-function of problem (4.1)–(4.2) associated with a given policy $\pi \in \Pi$ is defined as, for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$,*

$$(4.13) \quad q(y, \tilde{\theta}; \pi) := H\left(\tilde{\theta}, y, J_y(y; \pi), J_{yy}(y; \pi)\right) + \frac{\sigma_Z^2(y+1)^2}{2} J_{yy}(y; \pi) - \rho J(y; \pi).$$

Moreover, we notice that the difference between the q-function and the Hamiltonian is independent of θ , and this allows one to use the q-function for a policy improvement theorem. Thus, the policy improvement mapping \mathcal{I} in Theorem 4.1 can be expressed in terms of the q-function by $\mathcal{I}(\pi) = \frac{\exp\{\frac{1}{\gamma} q(y, \cdot; \pi)\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma} q(y, \tilde{\theta}; \pi)\} d\tilde{\theta}}$. This implies that the policy improvement iteration in Theorem 4.1 can be conducted by learning the q-function.

The following proposition gives the martingale condition to characterize the q-function for a given policy π when the value function is given.

PROPOSITION 4.2. *Let a policy $\pi \in \Pi$, its value function $J \in C^2(\mathbb{R}_+)$ satisfying (4.6), and a continuous function $\hat{q} : \mathbb{R}_+ \times \mathcal{A} \rightarrow \mathbb{R}$ be given. Then, $\hat{q}(y, \tilde{\theta}) = q(y, \tilde{\theta}; \pi)$ for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$ if and only if for all $y \in \mathbb{R}_+$, the following process,*

$$(4.14) \quad e^{-\rho t} J(Y_t^\pi; \pi) - \int_0^t e^{-\rho s} \hat{q}\left(Y_s^\pi, \tilde{\theta}_s^\pi\right) ds - \int_0^t e^{-\rho s} dL_s^\pi, \quad t \geq 0,$$

is an (\mathbb{F}, \mathbb{Q}) -martingale, where $Y^\pi = (Y_t^\pi)_{t \geq 0}$ is the solution to (4.1) with $Y_0^\pi = y$.

In the following, we strengthen Proposition 4.2 and characterize the q -function and the value function associated with a given policy π simultaneously. This result is the crucial theoretical tool for designing the q -learning algorithm.

THEOREM 4.3. *Let a policy $\pi \in \Pi$, a function $\hat{J} \in C^2(\mathbb{R}_+)$, and a continuous function $\hat{q}: \mathbb{R}_+ \times \mathcal{A} \rightarrow \mathbb{R}$ be given such that*

$$(4.15) \quad \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}} \left[e^{-\rho T} \hat{J}(Y_T^{\pi}) \right] = 0,$$

$$(4.16) \quad \int_{\mathcal{A}} [\hat{q}(y, \tilde{\theta}) - \gamma \ln \pi(\tilde{\theta} | y)] \pi(\tilde{\theta} | y) d\tilde{\theta} = 0 \quad \forall y \in \mathbb{R}_+.$$

Then, \hat{J} and \hat{q} are, respectively, the value function satisfying (4.6) and the q -function associated with π if and only if for all $y \in \mathbb{R}_+$, the following process,

$$e^{-\rho t} \hat{J}(Y_t^{\pi}) - \int_0^t e^{-\rho s} \hat{q}(Y_s^{\pi}, \tilde{\theta}_s^{\pi}) ds - \int_0^t e^{-\rho s} dL_s^{\pi}, \quad t \geq 0,$$

is an (\mathbb{F}, \mathbb{Q}) -martingale, where $Y^{\pi} = (Y_t^{\pi})_{t \geq 0}$ is the solution to (4.1) with $Y_0^{\pi} = y$. If it holds further that $\pi(\tilde{\theta} | y) = \frac{\exp\{\frac{1}{\gamma} \hat{q}(y, \tilde{\theta})\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma} \hat{q}(y, \tilde{\theta})\} d\tilde{\theta}}$ satisfying

$$(4.17) \quad \limsup_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}} \left[e^{-\rho T} \hat{J}(Y_T^{\pi'}) \right] \geq 0 \quad \forall \pi' \in \Pi,$$

then π is the optimal policy and \hat{J} is the optimal value function.

4.2. Continuous-time q -learning algorithm. In this subsection, we design q -learning algorithms to simultaneously learn and update the parameterized value function and the policy based on the martingale condition in Theorem 4.3.

Given a policy $\pi \in \Pi$, we parameterize the value function by a family of functions $J^{\xi}(\cdot)$, where $\xi \in \Theta \subset \mathbb{R}^{L_{\xi}}$ and L_{ξ} is the dimension of the parameter, and parameterize the q -function by a family of functions $q^{\psi}(\cdot, \cdot)$, where $\psi \in \Psi \subset \mathbb{R}^{L_{\psi}}$ and L_{ψ} is the dimension of the parameter. Moreover, we shall make the following assumptions for the parameterized family $\{J^{\xi}\}$ and $\{q^{\psi}\}$.

(A $_{\xi}$) The family of functions $J^{\xi}(\cdot)$ is C^1 in ξ and satisfies $\lim_{t \rightarrow \infty} \mathbb{E}^{\mathbb{Q}}[e^{-\rho t} J^{\xi}(Y_t^{\pi})] = 0$ and the Neumann condition $J_y^{\xi}(0) = 1$. Moreover, there exist continuous functions $G_J(\cdot)$ and $\tilde{J}(\cdot)$ such that $|J^{\xi}(y)| + |J_y^{\xi}(y)| + |J_{yy}^{\xi}(y)| \leq G_J(\xi) \tilde{J}(y)$ for all $(\xi, y) \in \Theta \times \mathbb{R}_+$.

(A $_{\psi}$) The family of functions $q^{\psi}(\cdot, \cdot)$ is C^1 in ψ and satisfies that

$$(4.18) \quad \int_{\mathcal{A}} [q^{\psi}(y, \tilde{\theta}) - \gamma \ln \pi(\tilde{\theta} | y)] \pi(\tilde{\theta} | y) d\tilde{\theta} = 0 \quad \forall y \in \mathbb{R}_+.$$

Furthermore, there exist continuous functions $G_q(\cdot)$ and $\tilde{q}(\cdot, \cdot)$ such that

$$(4.19) \quad |q^{\psi}(y, \tilde{\theta})| \leq G_q(\psi) \tilde{q}(y, \tilde{\theta}) \quad \forall (\psi, y, \tilde{\theta}) \in \Psi \times \mathbb{R}_+ \times \mathcal{A}.$$

Then, the learning task is to find the “optimal” (in some sense) parameters ξ and ψ . The key step in the algorithm design is to enforce the martingale condition stipulated in Theorem 4.3. More precisely, let $M = (M_t)_{t \geq 0}$ be the martingale given in Theorem 4.3, i.e.,

$$(4.20) \quad M_t = e^{-\rho t} J(Y_t^{\pi}) - \int_0^t e^{-\rho s} q(Y_s^{\pi}, \tilde{\theta}_s^{\pi}) ds - \int_0^t e^{-\rho s} dL_s^{\pi},$$

and $M^{\xi, \psi} = (M_t^{\xi, \psi})_{t \geq 0}$ be its parameterized process defined by

$$(4.21) \quad M_t^{\xi, \psi} = e^{-\rho t} J^{\xi}(Y_t^{\pi}) - \int_0^t e^{-\rho s} q^{\psi}(Y_s^{\pi}, \tilde{\theta}_s^{\pi}) ds - \int_0^t e^{-\rho s} dL_s^{\pi},$$

where $Y^{\pi} = (Y_t^{\pi})_{t \geq 0}$ is the solution to (4.1) with $Y_0^{\pi} = y$.

It follows from the martingale orthogonality condition that, for any test adapted continuous process $\varsigma = (\varsigma_t)_{t \geq 0}$ with $\mathbb{E}^{\mathbb{Q}}[\int_0^{\infty} |\varsigma_t|^2 d\langle M \rangle_t] < \infty$,

$$(4.22) \quad \mathbb{E}^{\mathbb{Q}} \left[\int_0^{\infty} \varsigma_t dM_t \right] = 0.$$

In fact, the following result shows that this is a necessary and sufficient condition for the parameterized process $M^{\xi, \psi}$ to be a martingale. Its proof is omitted because it is similar to the one of Proposition 4.2 in Jia and Zhou (2022b).

PROPOSITION 4.4. *The parameterized process $M^{\xi, \psi}$ given by (4.21) is a martingale if and only if*

$$(4.23) \quad \mathbb{E}^{\mathbb{Q}} \left[\int_0^{\infty} \varsigma_t dM_t^{\xi, \psi} \right] = 0$$

for any ς with $\mathbb{E}^{\mathbb{Q}}[\int_0^{\infty} |\varsigma_t|^2 d\langle M \rangle_t] < \infty$.

Proposition 4.4 tells us that, to find the optimal parameters ξ and ψ , it is enough to explore the solution (ξ^*, ψ^*) of the martingale orthogonality equation (4.23). This can be implemented by using stochastic approximation to update parameters as, for $\chi \in \{\xi, \psi\}$,

$$(4.24) \quad \chi \leftarrow \chi + \alpha_{\chi} \int_0^{\infty} \varsigma_t dM_t^{\xi, \psi},$$

where $\alpha_{\chi} > 0$ is the learning rate. However, the martingale orthogonality equation (4.23) and the associated update rule (4.24) may not be directly applicable due to its requirement of having full trajectories of the infinite horizon. To overcome this difficulty, we first truncate the martingale orthogonality equation at a sufficiently long time T :

$$(4.25) \quad \mathbb{E}^{\mathbb{Q}} \left[\int_0^T \varsigma_t dM_t^{\xi, \psi} \right] = 0.$$

Note that the truncated martingale orthogonality equation (4.25) involves integration in continuous time, which still cannot be directly applicable. Thus, we turn to consider the truncated discrete-time martingale orthogonality equation. Let $K \in \mathbb{N}$ and $\Delta t = T/K$, consider the partition $0 = t_0 < t_1 < t_2 < \dots < t_K = T$ with $t_k - t_{k-1} = \Delta t$ for $k = 1, \dots, K$. In light of (4.25), we present the truncated discretized martingale orthogonality equation as

$$(4.26) \quad \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \varsigma_{t_k} \left(M_{t_{k+1}}^{\xi, \psi} - M_{t_k}^{\xi, \psi} \right) \right] = 0.$$

The next theorem states the convergence of the stochastic approximation algorithms when $\Delta t \rightarrow 0$ and $T \rightarrow \infty$.

THEOREM 4.5. For a continuous test function ς satisfying $\mathbb{E}^{\mathbb{Q}}[\int_0^\infty |\varsigma_t|^2 d\langle M \rangle_t] < \infty$, consider the martingale orthogonality condition equations (4.23), (4.25), and (4.26). Then, we have

- (i) any convergent subsequence of the solution $(\xi_{\text{ML}}^*(T), \psi_{\text{ML}}^*(T))$ that solves the truncated martingale orthogonality equation (4.25) converges to the solution of the martingale orthogonality equation (4.23), that is,

$$\lim_{T \rightarrow \infty} (\xi_{\text{ML}}^*(T), \psi_{\text{ML}}^*(T)) = (\xi_{\text{ML}}^*, \psi_{\text{ML}}^*)$$

solves (4.23);

- (ii) any convergent subsequence of the solution $(\xi_{\text{ML}}^*(\Delta t; T), \psi_{\text{ML}}^*(\Delta t; T))$ that solves the truncated discretized martingale orthogonality equation (4.26) converges to the solution of the truncated martingale orthogonality equation (4.25), that is,

$$\lim_{\Delta t \rightarrow 0} (\xi_{\text{ML}}^*(\Delta t; T), \psi_{\text{ML}}^*(\Delta t; T)) = (\xi_{\text{ML}}^*(T), \psi_{\text{ML}}^*(T))$$

solves (4.25).

A direct result is that if the solution $(\xi_{\text{ML}}^*(\Delta t; T), \psi_{\text{ML}}^*(\Delta t; T))$ of the truncated discretized martingale orthogonality equation (4.26) converges, then it holds that

$$\lim_{T \rightarrow \infty} \lim_{\Delta t \rightarrow 0} (\xi_{\text{ML}}^*(\Delta t; T), \psi_{\text{ML}}^*(\Delta t; T)) = (\xi_{\text{ML}}^*, \psi_{\text{ML}}^*)$$

solves the martingale orthogonality equation (4.23). Therefore, it provides a theoretical foundation for implementing the truncation and discretization in the learning algorithm.

Next we use the truncated discretized martingale orthogonality equation (4.26) and Theorems 4.1, 4.3, and Theorem 4.5 to design the q-learning algorithm. To learn the optimal value function and q-function, we choose the proper parameterized function approximators J^ξ and q^ψ for $\xi \in \Theta \subset \mathbb{R}^{L_\xi}$, $\psi \in \Psi \subset \mathbb{R}^{L_\psi}$, which satisfy the assumptions (A_ξ) and (A_ψ) . Theorem 4.1 and the definition of the q-function give the parameterized form of the optimal policy as

$$(4.27) \quad \pi^\psi(\tilde{\theta} | y) = \frac{\exp\left\{\frac{1}{\gamma} q^\psi(y, \tilde{\theta})\right\}}{\int_{\mathcal{A}} \exp\left\{\frac{1}{\gamma} q^\psi(y, \tilde{\theta})\right\} d\tilde{\theta}} \quad \forall (y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}.$$

Moreover, it can be easily seen that with the policy π^ψ given in (4.27), the consistency condition in the assumption (A_ψ) trivially holds that

$$(4.28) \quad \int_{\mathcal{A}} \left[q^\psi(y, \tilde{\theta}) - \gamma \ln \pi^\psi(\tilde{\theta} | y) \right] \pi^\psi(\tilde{\theta} | y) d\tilde{\theta} = 0.$$

Based on the truncated discretized martingale orthogonality equation (4.26), we can apply the stochastic approximation method (Robbins and Monro (1951)) to update the parameters in the following manner, which corresponds to the TD(0) algorithm (cf. Sutton (1988) and Jia and Zhou (2022b)),

$$(4.29) \quad \xi \leftarrow \xi + \alpha_\xi \sum_{k=0}^{K-1} \iota_{t_k} G_k, \quad \psi \leftarrow \psi + \alpha_\psi \sum_{k=0}^{K-1} \varsigma_{t_k} G_k,$$

Algorithm 1. Offline q-learning algorithm.

Input: Initial state y_0 , horizon T , time step Δt , number of episodes N , number of mesh grids K , initial learning rates $\alpha_\xi(\cdot), \alpha_\psi(\cdot)$ (a function of the number of episodes), functional forms of parameterized value function $J^\xi(\cdot)$, q-function $q^\psi(\cdot)$, policy $\pi^\psi(\cdot | \cdot)$ and temperature parameter γ .

Required Program: an environment simulator $(y', L') = \text{Environment}_{\Delta t}(t, y, L, \tilde{\theta})$ that takes the current time-state pair (t, y, L) and action θ as inputs and generates state y' and L' at time $t + \Delta t$ as outputs.

Learning Procedure:

- 1: Initialize ξ, ψ , and $i = 1$.
- 2: **while** $i < N$ **do**
- 3: Initialize $j = 0$. Observe initial state y_0 and store $y_{t_0} \leftarrow y_0$.
- 4: **while** $j < K$ **do**
- 5: Generate action $\tilde{\theta}_{t_j} \sim \pi^\psi(\cdot | y_{t_j})$.
- 6: Apply $\tilde{\theta}_{t_j}$ to environment simulator $(y, L) = \text{Environment}_{\Delta t}(t_j, y_{t_j}, L_{t_j}, \tilde{\theta}_{t_j})$.
- 7: Observe new state y and L as output. Store $y_{t_{j+1}} \leftarrow y$, and $L_{t_{j+1}} \leftarrow L$.
- 8: **end while**
- 9: For every $k = 0, 1, \dots, K - 1$, compute

$$G_k = J^\xi(y_{t_{k+1}}) - J^\xi(y_{t_k}) - q^\psi(y_{t_k}, \tilde{\theta}_{t_k}) \Delta t - (L_{t_{k+1}} - L_{t_k}) - \rho J^\xi(y_{t_k}) \Delta t.$$
- 10: Update ξ and ψ by

$$\begin{aligned} \xi &\leftarrow \xi + \alpha_\xi(i) \sum_{k=0}^{K-1} e^{-\rho t_k} \frac{\partial J^\xi}{\partial \xi}(y_{t_k}) G_k, \\ \psi &\leftarrow \psi + \alpha_\psi(i) \sum_{k=0}^{K-1} e^{-\rho t_k} \frac{\partial q^\psi}{\partial \psi}(y_{t_k}, \tilde{\theta}_{t_k}) G_k. \end{aligned}$$

- 11: Update $i \leftarrow i + 1$.
 - 12: **end while**
-

where $\alpha_\xi, \alpha_\psi > 0$ are the learning rates, ι, ς are test functions, and for $k = 0, 1, \dots, K - 1$, the quantity G_k is given by

$$\begin{aligned} G_k &:= e^{-\rho t_{k+1}} J^\xi(Y_{t_{k+1}}^{\pi^\psi}) - e^{-\rho t_k} J^\xi(Y_{t_k}^{\pi^\psi}) - e^{-\rho t_k} q^\psi(Y_{t_k}^{\pi^\psi}, \theta_{t_k}^{\pi^\psi}) \Delta t \\ &\quad - e^{-\rho t_k} (L_{t_{k+1}}^{\pi^\psi} - L_{t_k}^{\pi^\psi}). \end{aligned}$$

Note that, by using the martingale orthogonality condition, we need at least $L_\xi + L_\psi$ equations to fully determine the parameters (ξ, ψ) . In this paper, we choose the test functions in the conventional sense that

$$\iota_t = e^{-\rho t} \frac{\partial J^\xi}{\partial \xi}(Y_t^{\pi^\psi}), \quad \varsigma_t = e^{-\rho t} \frac{\partial q^\psi}{\partial \psi}(Y_s^{\pi^\psi}, \theta_s^{\pi^\psi}).$$

Based on the above updating rules, we then present the pseudocode of the offline q-learning algorithm in Algorithm 1.

5. Numerical examples. In this section, we consider our auxiliary control problem with reflections as an example beyond the LQ control framework to illustrate the proposed q-learning algorithm. In particular, we will first choose the temperature parameter $\gamma = \rho/d$, where we recall that ρ is the discount factor and d is the number of risky assets. We also note that, in this case, Theorem 4.1(iii) asserts that the map \mathcal{I} has a fixed point π^* given by (4.12). However, it remains to verify the transversality condition (4.11) to check the optimality of π^* .

In what follows, we first give the explicit solution of the exploratory HJB equation (5.1) and show that π^* is indeed the optimal policy. Plugging (4.9) into the exploratory HJB equation (4.7), we get that

$$(5.1) \quad \frac{1}{2}\sigma_Z^2\kappa^2(y+1)^2v_{yy}(y) + \sqrt{1-\kappa^2}\zeta(y+1)v_y(y) + \gamma \ln \left(\sqrt{-\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|(v_{yy}(y))^d}} \right) - \alpha \frac{(v_y(y))^2}{v_{yy}(y)} = \rho v(y)$$

with the Neumann boundary condition $v_y(0) = 1$. We also recall that $\alpha = \frac{1}{2}\tilde{\mu}^\top(\sigma\sigma^\top)^{-1}\tilde{\mu}$ and $\zeta = \sigma_Z\eta^\top\sigma^{-1}\tilde{\mu}$. We conjecture that $v(y)$ for $y \geq 0$ satisfies the form of

$$(5.2) \quad v(y) = A \ln(1+y) + B$$

for some constants A and B . By plugging the expression into (5.1), (5.2) and using the choice of $\gamma = \rho/d$, we obtain that

$$A = 1, \quad B = \frac{1}{\rho} \left(-\frac{1}{2}\sigma_Z^2\kappa^2 + \sqrt{1-\kappa^2}\zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|} \right) + \alpha \right).$$

It then follows that

$$(5.3) \quad v(y) = \ln(1+y) + \frac{1}{\rho} \left(-\frac{1}{2}\sigma_Z^2\kappa^2 + \sqrt{1-\kappa^2}\zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|} \right) + \alpha \right)$$

is a classical solution of (5.1). For the general temperature parameter $\gamma \neq \rho/d$, the classical solution to the exploratory HJB equation (5.1), if it exists, is not in the simple form of (5.3) and actually does not admit an explicit expression, which shows the significant influence of the entropy regularizer in our stochastic control problem when it is not the LQ type control.

Back to the special choice of $\gamma = \rho/d$, as a result of (4.9) and (5.3), the candidate optimal policy is given by the following Gaussian policy,

$$(5.4) \quad \pi^*(\cdot|y) = \mathcal{N} \left(\tilde{\theta} | (\sigma\sigma^\top)^{-1}(\tilde{\mu} + \sigma_z\sqrt{1-\kappa^2}\sigma\eta)(1+y), (\sigma\sigma^\top)^{-1}\gamma(1+y)^2 \right),$$

which is also the fixed point as shown in Theorem 4.1(iii). It is notable that both the mean and the variance of the Gaussian policy $\pi^*(\cdot|y)$ depend on the state variable y and, in particular, its variance is increasing in y . That is, for the larger state process Y_t or the state process X_t , the agent needs to implement the more random policies with the larger variance for the purpose of learning.

The following verification theorem shows that the classical solution of the exploratory HJB equation (4.7) coincides with the value function and provides the optimal policy.

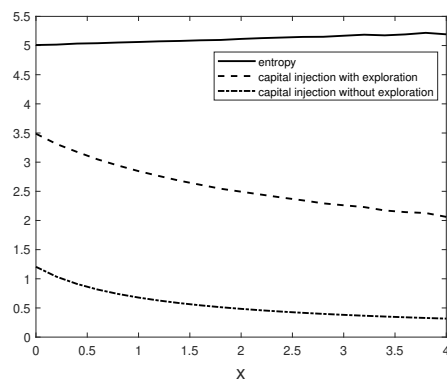


FIG. 1. Expected entropy and expected capital injection.

THEOREM 5.1 (verification theorem). *Considering the classical solution $v(y)$ for $y \in \mathbb{R}_+$ of the exploratory HJB equation (5.1) given by (5.3), the policy π^* given by (5.4), and the controlled state process $Y^* = (Y_t^*)_{t \geq 0}$ that obeys the following reflected SDE, for all $t \geq 0$,*

$$(5.5) \quad \begin{aligned} dY_t^* &= (2\alpha + \sqrt{1 - \kappa^2}\zeta)(1 + Y_t^*)dt \\ &+ \sqrt{\alpha + \frac{d}{2}\gamma + \frac{1}{2}\kappa^2\sigma_Z^2 + \sqrt{1 - \kappa^2}\zeta(1 + Y_t^*)}dB_t + dL_t^*. \end{aligned}$$

Here, $L^* = (L_t^*)_{t \geq 0}$ is the local time process for the process $Y^* = (Y_t^*)_{t \geq 0}$ at the reflecting boundary 0, namely, $L^* = (L_t^*)_{t \geq 0}$ is a continuous nondecreasing process satisfying $\int_0^t \mathbf{1}_{Y_s^*=0} dL_s^* = L_t^*$. Then, π^* is an optimal policy and $v(y)$ is the value function. That is, for all admissible $\pi \in \Pi$, we have $J(y; \pi) \leq v(y)$ for all $y \in \mathbb{R}_+$, where the equality holds when $\pi = \pi^*$.

Remark 5.2. When the model parameters are known, we present in Figure 1 the numerical comparison between the expectation of the discounted capital injection with exploration and the expectation of the discounted capital injection without exploration, illustrating the influence by the additional entropy regularizer or the policy exploration. As shown in Figure 1, both expectations of the discounted capital injection with exploration and without exploration are decreasing with respect to the variable y , which can be explained by the fact that the larger value of y indicates the larger initial distance between the portfolio process V^θ and the benchmark process Z , thereby it will have smaller chances for the capital injection. More importantly, Figure 1 shows that if the fund manager employs the exploratory policy to learn the optimal policy in the unknown environment, the randomized actions will inevitably cause the underlying controlled processes to hit the reflection boundary 0 more often compared with the case using the strict control, thereby leading to the higher expected capital injection.

Let us denote some parameters $(\xi^*, \psi_1^*, \psi_2^*, \psi_3^*)$ by

$$(5.6) \quad \begin{cases} \xi^* = \frac{1}{\rho} \left(-\frac{1}{2}\sigma_Z^2\kappa^2 + \sqrt{1 - \kappa^2}\zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|} \right) + \alpha \right), \\ \psi_1^* = \tilde{\mu} + \sigma_z \sqrt{1 - \kappa^2}\sigma\eta, \quad \psi_2^* = \sigma, \\ \psi_3^* = \frac{1}{2}\sigma_Z^2(\kappa^2 - 1) - \alpha - \sqrt{1 - \kappa^2}\zeta - \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|} \right). \end{cases}$$

Using Definition 4.2, the q-function can be expressed by, for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$,

$$(5.7) \quad q(y, \tilde{\theta}) = \frac{(\psi_1^*)^\top \tilde{\theta}}{1+y} - \frac{\tilde{\theta}^\top \psi_2^* (\psi_2^*)^\top \tilde{\theta}}{2(1+y)^2} - \rho \ln(1+y) + \psi_3^*,$$

and the value function $v(y)$ can be written

$$(5.8) \quad v(y) = \ln(1+y) + \xi^* \quad \forall y \in \mathbb{R}_+.$$

Based on (5.7)–(5.8), for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$, we can parameterize the optimal value function and the optimal q-function in the exact form as

$$(5.9) \quad \begin{cases} J^\xi(y) = \ln(1+y) + \xi, \\ q^\psi(y, \tilde{\theta}) = \frac{\psi_1^\top \tilde{\theta}}{1+y} - \frac{\tilde{\theta}^\top \psi_2 \psi_2^\top \tilde{\theta}}{2(1+y)^2} - \rho \ln(1+y) + \psi_3 \end{cases}$$

with the parameters $(\xi, \psi_1, \psi_2) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d \times d}$ to be learned, the parameter ψ_3 satisfying

$$\psi_3 = -\frac{1}{2} \psi_1^\top (\psi_2 \psi_2^\top)^{-1} \psi_1 - \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\psi_2 \psi_2^\top|} \right),$$

and the optimal policy can be parameterized by $\pi^\psi(\theta | y) = \frac{\exp\{\frac{1}{\gamma} q^\psi(y, \theta)\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma} q^\psi(y, \theta)\} d\theta}$. We can verify that the parameterized value function and q-function satisfy assumptions (A_ξ) and (A_ψ) .

We consider the model with one risky asset (i.e., $d = 1$), and we set the coefficients of the simulator to be $\tilde{\mu} = 0.2$, $\sigma = 1$, $\sigma_Z = 0.2$, and $\kappa = 0.5$. Furthermore, we set $\gamma = \rho = 0.2$ and the truncated horizon $T = 12$, the time step $\Delta t = 0.005$, and the number of episodes $N = 2 \times 10^4$. The learning rates $(\alpha_\psi, \alpha_\xi)$ are chosen by

$$\alpha_\psi(i) = \begin{cases} \left(\frac{0.1}{i^{0.61}}, \frac{0.01}{i^{0.61}} \right), & 1 \leq i \leq 1 \cdot 10^4, \\ \left(\frac{0.05}{i^{0.81}}, \frac{0.005}{i^{0.81}} \right), & 1 \cdot 10^4 < i \leq 2 \cdot 10^4, \end{cases}$$

$$\alpha_\xi(i) = \begin{cases} \frac{0.015}{i^{0.61}}, & 1 \leq i \leq 1 \cdot 10^4, \\ \frac{0.005}{i^{0.81}}, & 1 \cdot 10^4 < i \leq 2 \cdot 10^4, \end{cases}$$

which decay along the iterations. Based on Algorithm 1, we plot in Figure 2 the numerical results of the convergence of iterations for parameters $(\psi_1, \psi_2, \psi_3, \xi)$ in the optimal value function and the optimal q-function. The learned parameters for the optimal value function and the optimal q-function are summarized in Table 1.

For the general choice of temperature parameter $\gamma \neq \rho/d$, one needs to apply neural networks as the parameterized approximations of the optimal value function and the optimal q-function. In what follows, we implement our proposed q-learning algorithm together with neural networks. More precisely, we consider the parameterized value function $J^\xi(y)$ satisfying $J_y^\xi(0) = 1$ and transversality condition, and parameterized q-function given by $q^\psi(y, \tilde{\theta}) = -a^\psi(y)(\tilde{\theta} - b^\psi(y))^2 + c^\psi(y)$ with $c^\psi(y) = -\frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|b^\psi(y)|} \right)$ for $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$. Thus, we adopt three two-layer fully connected neural networks with sigmoid activation functions to train the functions $J^\xi(y)$ and $(a^\psi(y), b^\psi(y))$. Consider the model with one risky asset (i.e., $d = 1$), and still set

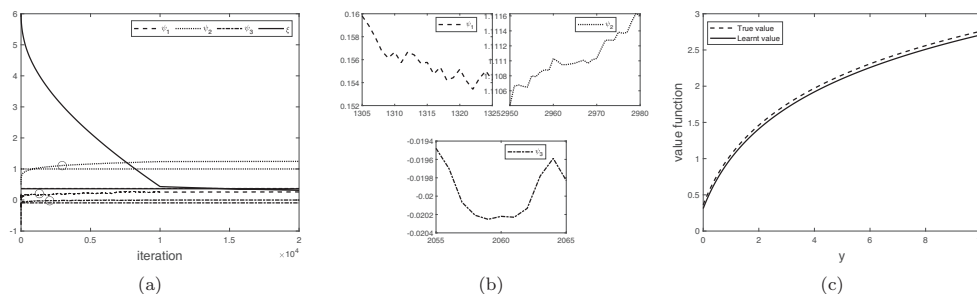


FIG. 2. Convergence of parameter iterations in the optimal value function and the optimal q -function using Algorithm 1. (a): paths of learned parameters for the optimal value function and the optimal q -function versus optimal parameters shown in the dashed line; (b): “zoom-in” paths of learned parameters for the hollow circles selected in panel (a) to illustrate the fluctuations of the local path; (c): the learned value function versus the optimal value function.

TABLE 1
The true value and learned value of the parameters $(\psi_1, \psi_2, \psi_3, \xi)$.

	ψ_1	ψ_2	ψ_3	ξ
True value	0.3732	1.0000	-0.0050	0.3624
Learned value	0.2520	1.2410	-0.0002	0.3326

the coefficients of the simulator to be $\tilde{\mu} = 0.2$, $\sigma = 1$, $\sigma_Z = 0.2$, $\kappa = 0.5$, and $\rho = 0.2$. Here, we choose different temperature parameters $\gamma = 0.05, 0.1, 0.3$. The truncated horizon is $T = 12$ and the time step is $\Delta t = 0.1$. The learning rates are given by $\alpha_\psi(i) = \alpha_\xi(i) = \frac{0.001}{i^{0.61}}$. We plot in Figure 3 the learned value function, the mean, and variance functions of the learned policy by using Algorithm 1 together with the neural networks after $N = 100$ iterations of policy update.

We next also compare the performance of the proposed q -learning algorithm with the classical stochastic control approach using the MLE method for model parameters. In particular, we implement two methods based on real market data. We choose the S&P 500 index as the benchmark process and select the Amazon.com Inc (AMZN) as the risky asset. We set the discount factor $\rho = 0.1$, time step $\Delta t = 1$ (day) and initial wealth $v = 3281.1$ (10 higher than the initial value of the benchmark process). The daily returns from January 1, 2000 through July 30, 2020 are used as the training set while daily returns from July 31, 2020 through July 30, 2024 are used as the test set.

For the MLE method, we first use the training set to obtain the estimated values of the parameters as $\hat{\mu} = 0.0012$, $\hat{\sigma} = 0.0324$, and $\hat{\sigma}_Z = 0.0126$. Suppose that the two BMs W^κ and W^η are independent, i.e., $\kappa = 1$. We then compute the optimal portfolio strategy given by (3.15) on the test set. On the other hand, for the reinforcement learning approach, we choose the temperature parameter $\gamma = \rho/d = 0.1$, and train $N = 500$ times with learning rates given by

$$\alpha_\psi(i) = \left(\frac{0.005}{i^{0.71}}, \frac{0.005}{i^{0.71}} \right), \quad \alpha_\xi = \frac{0.005}{i^{0.71}}, \quad 1 \leq i \leq 500,$$

on the training set. We then implement the q -learning algorithm on the test set with the learned value function and q -function. We plot the total wealth (including capital injection) process of the agent and the cumulative capital injection process in Figure 4. It shows that the wealth process using the q -learning algorithm outperforms the one using the MLE method. The total injected capital required by the agent using the reinforcement learning method is 2243.46, which is 5.86% lower than the total capital

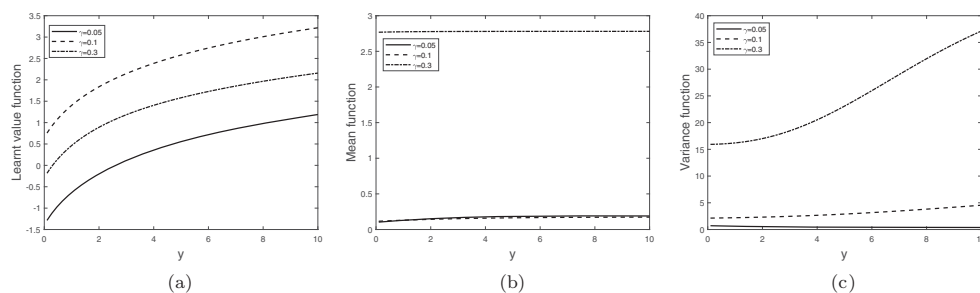


FIG. 3. The learned value function (left panel), mean function (middle panel), and variance function (right panel) of the learned policy by implementing Algorithm 1 together with the neural networks for $\gamma \neq \rho/d$.

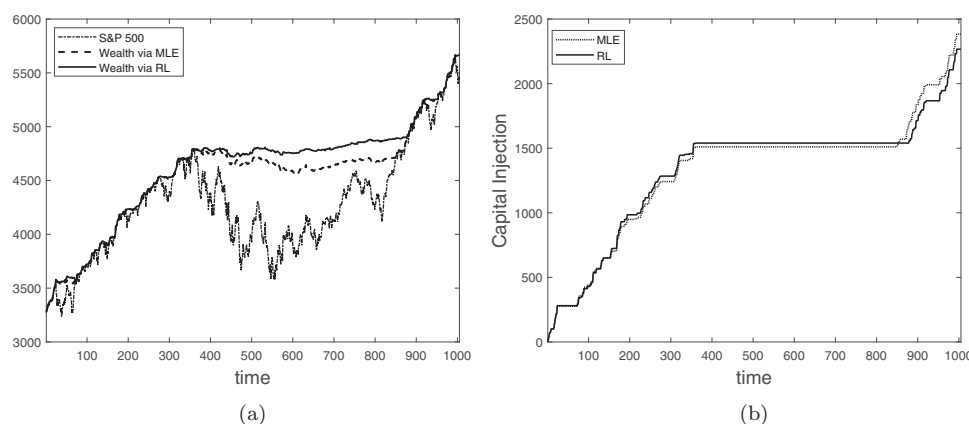


FIG. 4. (a): The price process of the S&P500, the total wealth (including capital injection) process under the MLE method and under the RL approach from July 31, 2020 through July 30, 2024. (b): The cumulative capital injection process under the MLE method and under the RL approach from July 31, 2020 through July 30, 2024.

injection of 2383.21 using the MLE method. These plots can show the effectiveness and robustness of the q-learning algorithm for some real market data.

6. Proofs. This section collects all proofs of the main results presented in previous sections.

Proof of Theorem 4.1. We first prove item (i). For $T > 0$ and $y \in \mathbb{R}_+$, applying Itô's formula to the process $e^{-\rho t} J(\tilde{Y}_t^{\pi'}; \pi)$ from 0 to T , we can obtain

$$\begin{aligned}
 & e^{-\rho T} J(\tilde{Y}_T^{\pi'}; \pi) - J(y; \pi) - \int_0^T e^{-\rho t} dL_t^{\pi'} - \gamma \int_0^T \int_{\mathcal{A}} e^{-\rho t} \ln \left(\pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) \right) \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) d\tilde{\theta} dt \\
 &= \int_0^T e^{-\rho t} \int_{\mathcal{A}} \left(\left[H \left(\tilde{\theta}, J_y(\tilde{Y}_t^{\pi'}; \pi), J_{yy}(\tilde{Y}_t^{\pi'}; \pi) \right) - \rho J(\tilde{Y}_t^{\pi'}; \pi) \right] \right. \\
 &\quad \left. + \frac{\sigma_Z^2 (\tilde{Y}_t^{\pi'} + 1)^2}{2} J_{yy}(\tilde{Y}_t^{\pi'}; \pi) \right. \\
 &\quad \left. - \gamma \ln \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) \right) \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) d\tilde{\theta} dt + \int_0^T e^{-\rho t} \left(J_y(\tilde{Y}_t^{\pi'}; \pi) - 1 \right) dL_t^{\pi'} \\
 & \quad + \int_0^T e^{-\rho t} \tilde{\sigma}(\tilde{Y}_t^{\pi'}, \pi'(\cdot | \tilde{Y}_t^{\pi'})) J_y(\tilde{Y}_t^{\pi'}; \pi) dB_t.
 \end{aligned}
 \tag{6.1}$$

As $J(\cdot; \pi) \in C^2(\mathbb{R}_+)$ satisfies the PDE (4.6), we can see that

$$(6.2) \quad \int_0^T e^{-\rho t} \left(J_y(\tilde{Y}_t^{\pi'}; \pi) - 1 \right) dL_t^{\pi'} = \int_0^T e^{-\rho t} (J_y(0; \pi) - 1) dL_t^{\pi'} = 0.$$

Moreover, it follows from Lemma 2 in Jia and Zhou (2023) that, for all $y \in \mathbb{R}_+$,

$$(6.3) \quad \begin{aligned} & \int_{\mathcal{A}} \left(\left[H(\tilde{\theta}, J_y(y; \pi), J_{yy}(y; \pi)) - \rho J(y; \pi) \right] \right. \\ & \quad \left. + \frac{\sigma_Z^2(y+1)^2}{2} J_{yy}(y; \pi) - \gamma \ln \pi'(\tilde{\theta}; y) \right) \pi'(\tilde{\theta}; y) d\tilde{\theta} \\ & \geq \int_{\mathcal{A}} \left(\left[H(\tilde{\theta}, J_y(y; \pi), J_{yy}(y; \pi)) - \rho J(y; \pi) \right] \right. \\ & \quad \left. + \frac{\sigma_Z^2(y+1)^2}{2} J_{yy}(y; \pi) - \gamma \ln \pi(\tilde{\theta}; y) \right) \pi(\tilde{\theta}; y) d\tilde{\theta} = 0. \end{aligned}$$

From (6.1), (6.2), and (6.3), we deduce that

$$(6.4) \quad \begin{aligned} & e^{-\rho T} J(\tilde{Y}_T^{\pi'}; \pi) - J(y; \pi) - \int_0^T e^{-\rho t} dL_t^{\pi'} \\ & - \gamma \int_0^T \int_{\mathcal{A}} e^{-\rho t} \ln \left(\pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) \right) \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) d\tilde{\theta} dt \\ & \geq \int_0^T e^{-\rho t} \tilde{\sigma}(\tilde{Y}_t^{\pi'}, \pi'(\cdot | \tilde{Y}_t^{\pi'})) J_y(\tilde{Y}_t^{\pi'}; \pi) dB_t. \end{aligned}$$

Taking the expectation on both side of (6.4) and letting $T \rightarrow \infty$, by (4.10) and the monotone convergence theorem (MCT), we conclude that for any $y \in \mathbb{R}_+$,

$$\begin{aligned} J(y; \pi) & \leq - \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[\int_0^T e^{-\rho t} dL_t^{\pi'} \right] \\ & \quad - \gamma \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[\int_0^T \int_{\mathcal{A}} e^{-\rho t} \ln \left(\pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) \right) \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) d\tilde{\theta} dt \right] \\ & \quad + \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} J(\tilde{Y}_T^{\pi'}; \pi) \right] \\ & = \mathbb{E}^{\mathbb{Q}^W} \left[- \int_0^\infty e^{-\rho t} dL_t^{\pi'} - \gamma \int_0^\infty \int_{\mathcal{A}} e^{-\rho t} \ln \left(\pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) \right) \pi'(\tilde{\theta}; \tilde{Y}_t^{\pi'}) d\tilde{\theta} dt \right] \\ & = J(y; \pi'). \end{aligned}$$

Next we deal with item (ii). Denote by $J^*(y) = J(y; \pi^*)$. Thus, $J^*(y)$ satisfies (4.6), and Lemma 2 in Jia and Zhou (2023) shows that

$$\begin{aligned} & \int_{\mathcal{A}} \left(\left[H(\tilde{\theta}, y, J_y^*(y), J_{yy}^*(y)) - \rho J^*(y) \right] \right. \\ & \quad \left. + \frac{\sigma_Z^2(y+1)^2}{2} J_{yy}^*(y) - \gamma \ln \pi^*(\tilde{\theta}; y) \right) \pi^*(\tilde{\theta}; y) d\tilde{\theta} \\ & = \sup_{\pi \in \mathcal{P}(\mathcal{A})} \int_{\mathcal{A}} \left(\left[H(\tilde{\theta}, y, J_y^*(y), J_{yy}^*(y)) - \rho J^*(y) \right] \right. \\ & \quad \left. + \frac{\sigma_Z^2(y+1)^2}{2} J_{yy}^*(y) - \gamma \ln \pi(\tilde{\theta}; y) \right) \pi(\tilde{\theta}; y) d\tilde{\theta} = 0. \end{aligned}$$

Fix $(T, y) \in (0, \infty) \times \mathbb{R}_+$ and $\pi \in \Pi$. By applying Itô's formula to $e^{-\rho T} J^*(\tilde{Y}_T^{\pi^*})$, we arrive at

$$(6.5) \quad \begin{aligned} e^{-\rho T} J^*(\tilde{Y}_T^{\pi}) &= J^*(y) + \int_0^T e^{-\rho s} J_y^*(\tilde{Y}_s^{\pi}) dL_s^Y + \int_0^T e^{-\rho s} (\mathcal{L}^{\pi}(\cdot | \tilde{Y}_s^{\pi}) J^* - \rho J^*)(\tilde{Y}_s^{\pi}) ds \\ &\quad + \int_0^T e^{-\rho s} J_y^*(\tilde{Y}_s^{\pi}) \tilde{\sigma}(\tilde{Y}_s^{\pi}, \pi(\cdot | \tilde{Y}_s^{\pi})) dB_s, \end{aligned}$$

where, for any $\pi \in \Pi$, the operator \mathcal{L}^{π} acting on $C^2(\mathbb{R}_+)$ is defined by

$$\mathcal{L}^{\pi} g(y) := \tilde{b}(y, \pi) g'(y) + \frac{1}{2} \tilde{\sigma}(y, \pi) g''(y), \quad \forall y \in \mathbb{R}_+.$$

Taking the expectation on both sides of (6.5), we deduce from the Neumann boundary condition $J_y^*(0) = 1$ that

$$(6.6) \quad \begin{aligned} &\mathbb{E}^{\mathbb{Q}^W} \left[- \int_0^T e^{-\rho s} dL_s^Y \right] \\ &= J^*(y) - \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} v(\tilde{Y}_T^{\pi}) \right] + \mathbb{E}^{\mathbb{Q}^W} \left[\int_0^T e^{-\rho s} (\mathcal{L}^{\pi}(\cdot | \tilde{Y}_s^{\pi}) v - \rho v)(\tilde{Y}_s^{\pi}) ds \right] \\ &\leq J^*(y) - \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} J^*(\tilde{Y}_T^{\pi}) \right]. \end{aligned}$$

Here, the last inequality in (6.6) holds true due to $(\mathcal{L}^{\pi} v - \rho v)(y) \leq 0$ for all $y \in \mathbb{R}_+$ and $\pi \in \mathcal{P}(\mathcal{A})$. Toward this end, letting $T \rightarrow \infty$ in (6.6), we obtain from (4.11) and MCT that, for all $\pi \in \Pi$,

$$\mathbb{E}^{\mathbb{Q}^W} \left[- \int_0^{\infty} e^{-\rho s} dL_s^Y \right] \leq v(y) \quad \forall y \in \mathbb{R}_+,$$

where the equality holds when $\pi = \pi^*$. This implies that $J^*(y) = v(y)$ is the optimal value function and π^* is the optimal policy.

Next, we move to item (iii). To find the fix point, we use the iteration method and start with a special Gaussian policy: $\pi_0(\cdot | y) \sim \mathcal{N}(\cdot | c_1(1+y), c_2(1+y)^2)$, with vector $c_1 \in \mathbb{R}^d$ and positive matrix $c_2 \in \mathbb{R}^{d \times d}$. Then it follows that the resulting value function $J^{\pi_0}(y)$ satisfies the equation

$$(6.7) \quad \begin{cases} \int_{\mathcal{A}} [H(\theta, y, J_y^{\pi_0}(y), J_{yy}^{\pi_0}(y)) - \gamma \log \pi_0(\theta | y)] \pi_0(\theta | y) d\theta + \frac{\sigma_Z^2}{2} (y+1)^2 J_{yy}^{\pi_0}(y) \\ \quad = \rho J^{\pi_0}(y), \\ J_y^{\pi_0}(y) = 1. \end{cases}$$

By standard arguments, we can show that the classical solution of (6.7) is given by

$$J^{\pi_0}(y) = \ln(1+y) + C \quad \forall y \in \mathbb{R}_+,$$

where C is a constant defined by

$$\begin{aligned} C := & \frac{1}{\rho} \left(\tilde{\mu}^\top c_1 + \sqrt{1 - \kappa^2} \sigma_Z c_1^\top \sigma \eta - \frac{1}{2} c_1^\top \sigma \sigma^\top c_1 - \frac{1}{2} \text{tr}(\sigma \sigma^\top c_2) \right. \\ & \left. - \frac{1}{2} \sigma_Z^2 + \frac{1}{2} \gamma \ln((2\pi)^d |c_2|) + \frac{d}{2} \gamma \right), \end{aligned}$$

and for a matrix $A \in \mathbb{R}^{d \times d}$, we denote $|A| := \det(A)$. Then, using once iteration, we get that

$$(6.8) \quad \pi_1(\cdot|y) = \mathcal{I}(\pi_0) \sim \mathcal{N}\left(\cdot \mid (\sigma\sigma^\top)^{-1}(\tilde{\mu} + \sigma_z \sqrt{1 - \kappa^2 \sigma \eta})(1+y), (\sigma\sigma^\top)^{-1}\gamma(1+y)^2\right).$$

Again, we can calculate the corresponding reward function as

$$(6.9) \quad J^{\pi_1}(y) = \ln(1+y) + \frac{1}{\rho} \left(-\frac{1}{2} \sigma_Z^2 \kappa^2 + \sqrt{1 - \kappa^2} \zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma\sigma^\top|} \right) + \alpha \right).$$

Then the iteration is applicable again, which yields the improved policy π_2 as exactly the Gaussian policy π_1 given in (6.8), together with the reward function in (6.9). Then we find that π^* give in (4.12) is a fixed point of \mathcal{I} . Thus, we complete the proof. \square

Proof of Proposition 4.2. For $T > 0$ and $y \in \mathbb{R}_+$, applying Itô's formula to the process $e^{-\rho s} J(Y_s^\pi; \pi)$ from 0 to t , we can obtain that

$$\begin{aligned} & e^{-\rho s} J(Y_s^\pi; \pi) - J(y; \pi) - \int_0^t e^{-\rho s} \hat{q}(Y_s^\pi, \tilde{\theta}^\pi) ds - \int_0^t e^{-\rho s} dL_s^\pi \\ &= \int_0^t e^{-\rho s} \left(H(\tilde{\theta}^\pi, Y_s^\pi, J_y(Y_s^\pi; \pi), J_{yy}(Y_s^\pi; \pi)) - \rho J(Y_s^\pi; \pi) \right. \\ & \quad \left. + \frac{\sigma_Z^2 (Y_s^\pi + 1)^2}{2} J_{yy}(Y_s^\pi; \pi) \right) ds \\ & \quad + \int_0^t e^{-\rho s} (J_y(Y_s^\pi; \pi) - 1) dL_s^\pi + \int_0^t e^{-\rho s} (\tilde{\theta}^\pi)^\top \sigma J_y(Y_s^\pi; \pi) d\tilde{W}_s \\ & \quad - \int_0^t e^{-\rho s} \sigma_Z (Y_s^\pi + 1) J_y(Y_s^\pi; \pi) d\tilde{W}_s^\kappa \\ &= \int_0^t e^{-\rho s} \left(q(Y_s^\pi, \tilde{\theta}^\pi; \pi) - \hat{q}(Y_s^\pi, \tilde{\theta}^\pi) \right) ds + \int_0^t e^{-\rho s} (\tilde{\theta}^\pi)^\top \sigma J_y(Y_s^\pi; \pi) d\tilde{W}_s \\ (6.10) \quad & - \int_0^t e^{-\rho s} \sigma_Z (Y_s^\pi + 1) J_y(Y_s^\pi; \pi) d\tilde{W}_s^\kappa. \end{aligned}$$

From (6.10), we can see that, if $\hat{q}(y, \tilde{\theta}) = q(y, \tilde{\theta}; \pi)$ for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$, then the above process and, hence, the process defined by (4.14) is an $((\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$ -martingale.

On the other hand, if the process (4.14) is an $((\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$ -martingale, we next show that $\hat{q}(y, \tilde{\theta}) = q(y, \tilde{\theta}; \pi)$ for $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$. By (6.10), $\int_0^t e^{-\rho s} (q(Y_s^\pi, \tilde{\theta}^\pi; \pi) - \hat{q}(Y_s^\pi, \tilde{\theta}^\pi)) ds$ for $t \geq 0$ is a continuous local martingale with finite variation and hence zero quadratic variation. Hence, it follows that $\int_0^t e^{-\rho s} (q(Y_s^\pi, \tilde{\theta}^\pi; \pi) - \hat{q}(Y_s^\pi, \tilde{\theta}^\pi)) ds = 0$ for all $t \in [0, \infty)$, \mathbb{Q} -a.s. (see Chapter 1, Exercise 5.21 in Karatzas and Shreve (2014)). For $y \in \mathbb{R}_+$, $a \in \mathcal{A}$, denote by $h(y, \tilde{\theta}) = q(y, \tilde{\theta}; \pi) - \hat{q}(y, \tilde{\theta})$, which is a continuous function that maps $\mathbb{R}_+ \times \mathcal{A}$ to \mathbb{R}_+ . Next, we argue by contradiction. Suppose there exists a pair $(y_0, \tilde{\theta}_0) \in \mathbb{R}_+ \times \mathcal{A}$ and $\epsilon > 0$ such that $h(y_0, \tilde{\theta}_0) > \epsilon$. By the continuity of h , there exists $\delta > 0$ such that $h(y, \tilde{\theta}) > \epsilon/2$ for all $(y, \tilde{\theta})$ with $\max\{|y - y_0|, |\tilde{\theta} - \tilde{\theta}_0|\} < \delta$.

Now, consider the state process, still denoted by Y^π , starting from $(y_0, \tilde{\theta}_0)$, namely, $(Y_t^\pi)_{t \geq 0}$ follows (6) with $Y_0^\pi = y_0$ and $\tilde{\theta}_0^\pi = \tilde{\theta}_0$. Define the stopping time τ by $\tau := \inf\{t \geq 0 : |Y_t^\pi - y_0| > \delta\}$. We have already shown that there exists $\Omega_0 \in \mathcal{F}$ with $\mathbb{Q}(\Omega_0) = 0$ such that for all $\omega \in \Omega \setminus \Omega_0$, $\int_0^t e^{-\rho s} (q(Y_s^\pi, \tilde{\theta}^\pi; \pi) - \hat{q}(Y_s^\pi, \tilde{\theta}^\pi)) ds = 0$ for all $t \in [0, \infty)$. It follows from Lebesgue's differentiation theorem that, for any

$\omega \in \Omega \setminus \Omega_0$, $h(Y_t^\pi(\omega), \tilde{\theta}_t^\pi(\omega)) = 0$, a.e., for $t \in [0, \tau(\omega)]$. Consider the set $\mathcal{O}(\omega) = \{t \in [0, \tau(\omega)] : |\tilde{\theta}_t^\pi(\omega) - \tilde{\theta}_0| < \delta\} \subset [0, \tau(\omega)]$. Because $h(Y_t^\pi(\omega), \tilde{\theta}_t^\pi(\omega)) > \frac{\epsilon}{2}$ when $t \in \mathcal{O}(\omega)$, we conclude that $\mathcal{O}(\omega)$ has Lebesgue measure zero for any $\omega \in \Omega \setminus \Omega_0$. That is

$$\int_0^\infty \mathbf{1}_{\{t \leq \tau(\omega)\}} \mathbf{1}_{\{|\tilde{\theta}_t^\pi(\omega) - \tilde{\theta}_0| < \delta\}} dt = 0.$$

Integrating ω with respect to \mathbb{Q} and applying Fubini's theorem, we obtain that

$$\begin{aligned} 0 &= \int_\Omega \int_0^\infty \mathbf{1}_{\{t \leq \tau(\omega)\}} \mathbf{1}_{\{|\tilde{\theta}_t^\pi(\omega) - \tilde{\theta}_0| < \delta\}} dt \mathbb{Q}(d\omega) \int_0^\infty \mathbb{E}^\mathbb{Q} \left[\mathbf{1}_{\{t \leq \tau\}} \mathbf{1}_{\{|\tilde{\theta}_t^\pi - \tilde{\theta}_0| < \delta\}} \right] dt \\ &= \int_0^\infty \mathbb{E} \left[\mathbf{1}_{\{t \leq \tau\}} \mathbb{Q}^\mathbb{Q} \left(|\tilde{\theta}_t^\pi - \tilde{\theta}_0| < \delta \mid \mathcal{F}_t \right) \right] dt = \int_0^\infty \mathbb{E} \left[\mathbf{1}_{\{t \leq \tau\}} \int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | Y_t^\pi) d\tilde{\theta} \right] dt \\ &\geq \min_{|y - y_0| < \delta} \left\{ \int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | y) d\tilde{\theta} \right\} \int_0^\infty \mathbb{E}^\mathbb{Q} [\mathbf{1}_{\{t \leq \tau\}}] dt \\ &= \min_{|y - y_0| < \delta} \left\{ \int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | y) d\tilde{\theta} \right\} \mathbb{E}^\mathbb{Q}[\tau] \geq 0. \end{aligned}$$

The above implies that $\min_{|y - y_0| < \delta} \int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | y) d\tilde{\theta} = 0$. However, this contradicts Definition 4.1 about an admissible policy. Indeed, Definition 4.1(i) stipulates $\text{supp } \pi(\cdot | y) = \mathcal{A}$ for any $y \in \mathbb{R}_+$; hence $\int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | y) d\tilde{\theta} > 0$. Then, the continuity in Definition 4.1(iii) yields $\min_{|y - y_0| < \delta} \int_{|\tilde{\theta} - \tilde{\theta}_0| < \delta} \pi(\tilde{\theta} | y) d\tilde{\theta} > 0$, which is a contradiction. Hence, we conclude that $q(y, \tilde{\theta}; \pi) = \hat{q}(y, \tilde{\theta})$ for every $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$. \square

Proof of Theorem 4.3. If \hat{J} and \hat{q} are, respectively, the value function and the q -function associated with the policy π , it follows from the same argument as in the proof of Proposition 4.2 that

$$e^{-\rho t} \hat{J}(Y_t^\pi) - \int_0^t e^{-\rho s} \hat{q}(Y_s^\pi, \tilde{\theta}_s^\pi) ds - \int_0^t e^{-\rho s} dL_s^\pi, \quad t \geq 0,$$

is an $((\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$ -martingale. On the other hand, assume that $e^{-\rho t} \hat{J}(Y_t^\pi) - \int_0^t e^{-\rho s} \hat{q}(Y_s^\pi, \tilde{\theta}_s^\pi) ds - \int_0^t e^{-\rho s} dL_s^\pi$ is an $((\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$ martingale. It then holds that, for any $T > 0$,

$$(6.11) \quad \hat{J}(y) = \mathbb{E}^\mathbb{Q} \left[e^{-\rho T} \hat{J}(Y_T^\pi) - \int_0^T e^{-\rho s} \hat{q}(Y_s^\pi, \tilde{\theta}_s^\pi) ds - \int_0^T e^{-\rho s} dL_s^\pi \right].$$

Integrating over the action randomization with respect to the policy π on the both sides of (6.11), we get that

$$(6.12) \quad \hat{J}(y) = \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} \hat{J}(\tilde{Y}_T^\pi; \pi) - \int_0^T e^{-\rho s} \int_{\mathcal{A}} \hat{q}(\tilde{Y}_s^\pi, \tilde{\theta}) \pi(\tilde{\theta} | \tilde{Y}_s^\pi) d\tilde{\theta} ds - \int_0^T e^{-\rho s} dL_s^\pi \right].$$

In view of that $\int_{\mathcal{A}} [\hat{q}(y, \tilde{\theta}) - \gamma \ln \pi(\tilde{\theta} | y)] \pi(\tilde{\theta} | y) d\tilde{\theta} = 0$, we obtain

$$(6.13) \quad \hat{J}(y) = \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} \hat{J}(\tilde{Y}_T^\pi) - \gamma \int_0^T e^{-\rho s} \int_{\mathcal{A}} \ln \pi(\tilde{\theta} | \tilde{Y}_s^\pi) \pi(\tilde{\theta} | \tilde{Y}_s^\pi) d\tilde{\theta} ds - \int_0^T e^{-\rho s} dL_s^\pi \right].$$

Letting T go to infinity on both side of (6.13), we deduce from MCT and (4.15) that

$$(6.14) \quad \hat{J}(y) = \mathbb{E}^{\mathbb{Q}^W} \left[-\gamma \int_0^\infty e^{-\rho s} \int_{\mathcal{A}} \ln \pi(\tilde{\theta} | \tilde{Y}_s^\pi) \pi(\tilde{\theta} | \tilde{Y}_s^\pi) d\tilde{\theta} ds - \int_0^\infty e^{-\rho s} dL_s^\pi \right].$$

This yields that $\hat{J}(y) = J(y; \pi)$ for all $y \in \mathbb{R}_+$ by virtue of (4.6). Furthermore, based on Proposition 4.2, the martingale condition implies that $\hat{q}(y, \tilde{\theta}) = q(y, \tilde{\theta}; \pi)$ for all $(y, \tilde{\theta}) \in \mathbb{R}_+ \times \mathcal{A}$.

Finally, if $\pi(\tilde{\theta} | y) = \frac{\exp\{\frac{1}{\gamma} \hat{q}(y, \tilde{\theta})\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma} \hat{q}(y, \tilde{\theta})\} d\tilde{\theta}}$ satisfies the condition (4.17), then $\pi = \mathcal{I}(\pi)$, where \mathcal{I} is the map defined in Theorem 4.1. This in turn implies that π is the optimal policy and \hat{J} is the optimal value function. \square

Proof of Theorem 4.5. We first state the following auxiliary lemma.

LEMMA 6.1 (Lemma 8 in Jia and Zhou (2022b)). *Let $f_h(x) = f(x) + r_h(x)$, where f is a continuous function and r_h converges to 0 uniformly on any compact set as $h \rightarrow 0$. Assume that $f_h(x_h^*) = 0$ and $\lim_{h \rightarrow 0} x_h^* = x^*$. Then $f(x^*) = 0$.*

We first apply Lemma 6.1 to prove item (i). We can take $f(x) = \mathbb{E}^{\mathbb{Q}}[\int_0^\infty \varsigma_t dM_t^{\xi, \psi}]$ and $f_h(x) = \mathbb{E}^{\mathbb{Q}}[\int_0^T \varsigma_t dM_t^{\xi, \psi}]$ with $h = 1/T$, $x = (\xi, \psi)$. Thus, we need to prove that

$$(6.15) \quad \mathbb{E}^{\mathbb{Q}} \left[\int_0^\infty \varsigma_t dM_t^{\xi, \psi} \right] - \mathbb{E}^{\mathbb{Q}} \left[\int_0^T \varsigma_t dM_t^{\xi, \psi} \right] = \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty \varsigma_t dM_t^{\xi, \psi} \right] \rightarrow 0,$$

as $T \rightarrow \infty$ uniformly on any compact subset of $\Theta \times \Psi$. Let $\mathcal{C} \subset \Theta \times \Psi$ be a compact set, then \mathcal{C} is a bounded closed set. By Proposition 4.4 and assumptions (A_ξ) and (A_ψ) , we have

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty \varsigma_t dM_t^{\xi, \psi} \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t \left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right) dt \right] + \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t dL_t^\pi \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t \left((\tilde{\theta}_t^\pi)^\top \tilde{\mu} J_y^\xi(Y_t^\pi) + \frac{1}{2} (\tilde{\theta}_t^\pi)^\top \sigma \sigma^\top \tilde{\theta}_t^\pi J_{yy}^\xi(Y_t^\pi) \right. \right. \\ &\quad \left. \left. - \sqrt{1 - \kappa^2} \sigma_Z(Y_t^\pi + 1) (\tilde{\theta}_t^\pi)^\top \sigma \eta J_{yy}^\xi(Y_t^\pi) \right. \right. \\ &\quad \left. \left. + \frac{1}{2} \sigma_Z^2(Y_t^\pi + 1)^2 J_{yy}^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right) dt \right] + \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t dL_t^\pi \right] \\ &\leq (G_J(\xi) + G_q(\psi)) \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t \left(\left((\tilde{\theta}_t^\pi)^\top \tilde{\mu} \right) + \frac{1}{2} (\tilde{\theta}_t^\pi)^\top \sigma \sigma^\top \tilde{\theta}_t^\pi \right. \right. \\ &\quad \left. \left. + \left| \sqrt{1 - \kappa^2} \sigma_Z(Y_t^\pi + 1) (\tilde{\theta}_t^\pi)^\top \sigma \eta \right| \right. \right. \\ (6.16) \quad &\quad \left. \left. + \frac{1}{2} \sigma_Z^2(Y_t^\pi + 1)^2 + \rho \right) \tilde{J}(Y_t^\pi) + \tilde{q}(Y_t^\pi, \tilde{\theta}_t^\pi) \right) dt \right] + \mathbb{E}^{\mathbb{Q}} \left[\int_T^\infty e^{-\rho t} \varsigma_t dL_t^\pi \right]. \end{aligned}$$

This yields that as $T \rightarrow \infty$, $\mathbb{E}^{\mathbb{Q}}[\int_T^\infty \varsigma_t dM_t^{\xi, \psi}]$ converges to 0 uniformly in $(\xi, \psi) \in \mathcal{C}$. By Lemma 6.1, we get the desired result.

Next we deal with item (ii). By Lemma 6.1 again, it is sufficient to show that

$$(6.17) \quad \mathbb{E}^{\mathbb{Q}} \left[\int_0^\infty \varsigma_t dM_t^{\xi, \psi} \right] - \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \varsigma_{t_k} \left(M_{t_{k+1}}^{\xi, \psi} - M_{t_k}^{\xi, \psi} \right) \right] \rightarrow 0,$$

as $\Delta t \rightarrow \infty$ uniformly on any compact subset of $\Theta \times \Psi$. Let $\mathcal{C} \subset \Theta \times \Psi$ be a compact set, which is also a bounded closed set. It follows that

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}} \left[\int_0^\infty \varsigma_t dM_t^{\xi, \psi} \right] - \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \varsigma_{t_k} \left(M_{t_{k+1}}^{\xi, \psi} - M_{t_k}^{\xi, \psi} \right) \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\int_0^\infty \varsigma_t dM_t^{\xi, \psi} \right] - \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \varsigma_{t_k} dM_t^{\xi, \psi} \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\rho t} (\varsigma_t - \varsigma_{t_k}) \left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right) dt \right] \\ &\quad + \mathbb{E}^{\mathbb{Q}} \left[\int_0^T e^{-\rho t} \varsigma_t dL_t^\pi - \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\rho t} \varsigma_{t_k} dL_t^\pi \right]. \end{aligned}$$

Note that

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}} \left[\int_0^T e^{-\rho t} \varsigma_t dL_t^\pi - \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\rho t} \varsigma_{t_k} dL_t^\pi \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\int_0^T e^{-\rho t} \varsigma_t dL_t^\pi - \sum_{k=0}^{K-1} e^{-\rho t_k} \varsigma_{t_k} \left(L_{t_{k+1}}^\pi - L_{t_k}^\pi \right) \right] \\ &\quad + \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} (e^{-\rho t} - e^{-\rho t_k}) \varsigma_{t_k} dL_t^\pi \right] \\ &\leq \mathbb{E}^{\mathbb{Q}} \left[\int_0^T e^{-\rho t} \varsigma_t dL_t^\pi - \sum_{k=0}^{K-1} e^{-\rho t_k} \varsigma_{t_k} \left(L_{t_{k+1}}^\pi - L_{t_k}^\pi \right) \right] \\ &\quad + \sup_{|t-s| \leq \Delta t, t, s \in [0, T]} (e^{-\rho t} - e^{-\rho s}) \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \varsigma_{t_k} dL_t^\pi \right], \end{aligned}$$

where the first term in the above inequality converges to 0 as $\Delta t \rightarrow 0$ by using the definition of the Riemann–Stieltjes integral and the second term converges to 0 as $\Delta t \rightarrow 0$ since $\sup_{|t-s| \leq \Delta t, t, s \in [0, T]} (e^{-\rho t} - e^{-\rho s}) \rightarrow 0$. On the other hand, we have

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\rho t} (\varsigma_t - \varsigma_{t_k}) \left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right) dt \right] \\ &\leq \sum_{k=0}^{K-1} \left(\int_{t_k}^{t_{k+1}} \mathbb{E}^{\mathbb{Q}} [(\varsigma_t - \varsigma_{t_k})^2] dt \right)^{\frac{1}{2}} \\ &\quad \times \left(\int_{t_k}^{t_{k+1}} \mathbb{E}^{\mathbb{Q}} \left[\left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right)^2 \right] dt \right)^{\frac{1}{2}} \\ &\leq d(\ell(\cdot), \Delta t) \sqrt{\Delta t} \sum_{k=0}^{K-1} \left(\int_{t_k}^{t_{k+1}} \mathbb{E}^{\mathbb{Q}} \left[\left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right)^2 \right] dt \right)^{\frac{1}{2}} \\ &\leq d(\ell(\cdot), \Delta t) \sqrt{\Delta t} \sqrt{K} \left(\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}^{\mathbb{Q}} \left[\left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right)^2 \right] dt \right)^{\frac{1}{2}} \end{aligned}$$

$$\begin{aligned}
&\leq d(\ell(\cdot), \Delta t) \sqrt{T} \left(\mathbb{E}^{\mathbb{Q}} \left[\int_0^T \left(\mathcal{L}^{\tilde{\theta}_t^\pi} J^\xi(Y_t^\pi) - \rho J^\xi(Y_t^\pi) - q^\psi(Y_t^\pi, \tilde{\theta}_t^\pi) \right)^2 dt \right] \right)^{\frac{1}{2}} \\
&\leq d(\ell(\cdot), \Delta t) \sqrt{T} (G_J^2(\xi) + G_q^2(\psi))^{\frac{1}{2}} \mathbb{E}^{\mathbb{Q}} \left[\int_0^T \left(\left| (\tilde{\theta}_t^\pi)^\top \tilde{\mu} \right| + \frac{1}{2} (\tilde{\theta}_t^\pi)^\top \sigma \sigma^\top \tilde{\theta}_t^\pi \right. \right. \\
&\quad \left. \left. + \left| \sqrt{1 - \kappa^2} \sigma_Z (Y_t^\pi + 1) (\tilde{\theta}_t^\pi)^\top \sigma \eta \right| + \frac{1}{2} \sigma_Z^2 (Y_t^\pi + 1)^2 + \rho \right) \tilde{J}(Y_t^\pi) + \tilde{q}(Y_t^\pi, \tilde{\theta}_t^\pi) \right)^2 dt \right]^{\frac{1}{2}},
\end{aligned}$$

where $\ell(t) := \varsigma_t$ and $d(\ell(\cdot), \Delta t) = \sup_{|t-s| \leq \Delta t, t, s \in [0, T]} \mathbb{E}^{\mathbb{Q}}[|\ell(t) - \ell(s)|^2]^{\frac{1}{2}}$ is the modulus of continuity of $\ell(\cdot)$ in $\mathbb{L}^2(\Omega)$. Therefore, $d^2(\ell(\cdot), \Delta t) \rightarrow 0$ as $\Delta t \rightarrow 0$ and we get the desired result (6.17). \square

Proof of Theorem 5.1. By Theorem 4.1, it suffices to verify the transversality conditions in (4.11). Noting that $y \rightarrow v(y)$ is nondecreasing, we have

$$\begin{aligned}
v(y) &= \ln(1 + y) + \frac{1}{\rho} \left(-\frac{1}{2} \sigma_Z^2 \kappa^2 + \sqrt{1 - \kappa^2} \zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma \sigma^\top|} \right) + \alpha \right) \\
&\geq \frac{1}{\rho} \left(-\frac{1}{2} \sigma_Z^2 \kappa^2 + \sqrt{1 - \kappa^2} \zeta + \frac{\gamma}{2} \ln \left(\frac{(2\pi\gamma)^d}{|\sigma \sigma^\top|} \right) + \alpha \right).
\end{aligned}$$

This yields that, for any $\pi \in \Pi$,

$$(6.18) \quad \limsup_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} v(Y_T^\pi) \right] \geq 0.$$

Next, we check the validity of the second transversality condition in (4.11), that is,

$$(6.19) \quad \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{Q}^W} \left[e^{-\rho T} v(Y_T^*) \right] = 0.$$

It follows from (4.3) and (5.3) that the controlled state process $Y^* = (Y_t^*)_{t \geq 0}$ with policy π^* obeys the following reflected SDE; for $t > 0$,

$$dY_t^* = (2\alpha + \sqrt{1 - \kappa^2} \zeta)(1 + Y_t^*) dt + \sqrt{\alpha + \frac{d}{2} \gamma + \frac{1}{2} \kappa^2 \sigma_Z^2 + \sqrt{1 - \kappa^2} \zeta (1 + Y_t^*)} dB_t + dL_t^*.$$

Introduce the processes $H_t := \ln(1 + Y_t^*)$ and $K_t := \int_0^t \frac{dL_s^*}{1 + Y_s^*}$ for $t \geq 0$. Then, $t \rightarrow H_t$ is a nonnegative process with $H_0 = h = \ln(1 + y)$. Moreover, $t \rightarrow K_t$ is a nondecreasing process and satisfies $\int_0^t \mathbf{1}_{H_s=0} dK_s = \int_0^t \mathbf{1}_{Y_s^*=0} \frac{dL_s^*}{1 + Y_s^*} = \int_0^t \frac{dL_s^*}{1 + Y_s^*} = K_t$ a.s. for $t \geq 0$. This implies that $K = (K_t)_{t \geq 0}$ is the local time of the process $H = (H_t)_{t \geq 0}$ at the reflecting boundary 0. By applying Itô's formula to $H_t = \ln(1 + Y_t^*)$, we arrive at

$$(6.20) \quad dH_t = \hat{\mu} dt + \hat{\sigma} dB_t + \frac{dL_t^*}{1 + Y_t^*} = \hat{\mu} dt + \hat{\sigma} dB_t + dK_t, \quad t > 0,$$

where constants $\hat{\mu}$ and $\hat{\sigma}$ are defined by

$$\begin{aligned}
\hat{\mu} &:= 2\alpha + \sqrt{1 - \kappa^2} \zeta - \frac{1}{2} \left(\alpha + \frac{d}{2} \gamma + \frac{1}{2} \kappa^2 \sigma_Z^2 + \sqrt{1 - \kappa^2} \zeta \right), \\
\hat{\sigma} &:= \sqrt{\alpha + \frac{d}{2} \gamma + \frac{1}{2} \kappa^2 \sigma_Z^2 + \sqrt{1 - \kappa^2} \zeta}.
\end{aligned}$$

From the solution representation of the “Skorokhod problem,” it follows that, for any $y \geq 0$,

$$(6.21) \quad K_t = 0 \vee \left\{ -h + \max_{s \in [0, t]} (-\hat{\mu}s - \hat{\sigma}B_s) \right\} \leq h + |\hat{\mu}|t + |\hat{\sigma}| \max_{s \in [0, t]} B_s.$$

Thus, we deduce from (6.20) and (6.21) that, for $T > 0$,

$$e^{-\rho T} \mathbb{E}^{\mathbb{Q}^W} [\ln(1 + Y_T^*)] = e^{-\rho T} \mathbb{E}^{\mathbb{Q}^W} [K_T] \leq e^{-\rho T} \left(2h + 2|\hat{\mu}|T + |\hat{\sigma}| \sqrt{\frac{2T}{\pi}} \right) \rightarrow 0, \text{ as } T \rightarrow \infty.$$

This yields the desired transversality condition (6.19) and completes the proof. \square

Acknowledgment. The authors are grateful to two anonymous referees for their helpful comments and suggestions.

REFERENCES

- L. BAI, T. GAMAGE, J. MA, AND P. XIE (2023), *Reinforcement Learning for Optimal Dividend Problem under Diffusion Model*, preprint, arXiv:2309.10242.
- T. BJÖRK AND C. LANDÉN (2000), *On the term structure of futures and forward prices*, in Mathematical Finance – Bachelier Congress 2000, H. Geman, D. Madan, S. R. Pliska, and T. Vorst, eds., Springer, Cham, Switzerland, pp. 111–149.
- L. BO, Y. HUANG, AND X. YU (2023), *An Extended Merton Problem with Relaxed Benchmark Tracking*, preprint, arXiv:2304.10802.
- L. BO, Y. HUANG, AND X. YU (2024), *Stochastic control problems with state-reflections arising from relaxed benchmark tracking*, Math. Oper. Res., to appear, <https://doi.org/10.1287/moor.2023.0265>.
- L. BO, H. LIAO, AND X. YU (2021), *Optimal tracking portfolio with a ratcheting capital benchmark*, SIAM J. Control Optim., 59, pp. 2346–2380, <https://doi.org/10.1137/20M1348856>.
- S. BROWNE (1999a), *Reaching goals by a deadline: Digital options and continuous-time active portfolio management*, Adv. Appl. Probab., 31, pp. 551–577, <https://doi.org/10.1239/aap/1029955147>.
- S. BROWNE (1999b), *Beating a moving target: Optimal portfolio strategies for outperforming a stochastic benchmark*, Finance Stoch., 3, pp. 275–294, <https://doi.org/10.1007/s007800050063>.
- S. BROWNE (2000), *Risk-constrained dynamic active portfolio management*, Manag. Sci., 46, pp. 1188–1199, <https://doi.org/10.1287/mnsc.46.9.1188.12233>.
- G. BRUNICK AND S. SHREVE (2013), *Mimicking an Itô process by a solution of a stochastic differential equation*, Ann. Appl. Probab., 23, pp. 1584–1628, <https://doi.org/10.1214/12-AAP881>.
- A. GAIIVORONSKI, S. KRYLOV, AND N. WIJST (2005), *Optimal portfolio selection and dynamic benchmark tracking*, European J. Oper. Res., 163, pp. 115–131, <https://doi.org/10.1016/j.ejor.2003.12.001>.
- Y. JIA AND X. Y. ZHOU (2022a), *Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms*, J. Mach. Learn. Res., 23, pp. 1–50.
- Y. JIA AND X. Y. ZHOU (2022b), *Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach*, J. Mach. Learn. Res., 23, pp. 1–55.
- Y. JIA AND X. Y. ZHOU (2023), *q-Learning in continuous time*, J. Mach. Learn. Res., 24, pp. 1–61.
- I. KARATZAS AND S. SHREVE (2014), *Brownian Motion and Stochastic Calculus*, Grad. Texts in Math. 113, Springer, New York.
- H. J. KUSHNER (1998), *Existence of optimal controls for variance control*, in Stochastic Analysis, Control, Optimization and Applications: A volume in honor of W.H. Fleming, W. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser, Boston.
- H. J. KUSHNER AND P. DUPUIS (1991), *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer, New York.
- C. NI, Y. LI, P. FORSYTH, AND R. CARROLL (2022), *Optimal asset allocation for outperforming a stochastic benchmark target*, Quant. Finance, 22, pp. 1595–1626, <https://doi.org/10.1080/14697688.2022.2072233>.
- H. ROBBINS AND S. MONRO (1951), *A stochastic approximation method*, Ann. Math. Statist., 22, pp. 400–407, <https://doi.org/10.1214/aoms/1177729586>.

- O. STRUB AND P. BAUMANN (2018), *Optimal construction and rebalancing of index-tracking portfolios*, European J. Oper. Res., 264, pp. 370–387, <https://doi.org/10.1016/j.ejor.2017.06.055>.
- Y. SUN (2006), *The exact law of large numbers via Fubini extension and characterization of insurable risks*, J. Econom. Theory, 126, pp. 31–69, <https://doi.org/10.1016/j.jet.2004.10.005>.
- R. S. SUTTON (1988), *Learning to predict by the methods of temporal differences*, Mach. Learn., 3, pp. 9–44, <https://doi.org/10.1007/BF00115009>.
- W. TANG, Y. P. ZHANG, AND X. Y. ZHOU (2022), *Exploratory HJB equation and their convergence*, SIAM J. Contr. Optim., 60, pp. 3191–3216, <https://doi.org/10.1137/21M1448185>.
- B. WANG, X. GAO, AND L. LI (2023), *Reinforcement Learning for Continuous-time Optimal Execution: Actor-Critic Algorithm and Error Analysis*, preprint, SSRN: 4378950.
- H. WANG, T. ZARIPHPOULOU, AND X. Y. ZHOU (2020), *Reinforcement learning in continuous time and space: A stochastic control approach*, J. Mach. Learn. Res., 21, pp. 1–34.
- X. WEI AND X. YU (2025), *Continuous-time Q-learning for mean-field control problems*, Appl. Math. Optim., 91.
- D. YAO, S. ZHANG, AND X. Y. ZHOU (2006), *Tracking a financial benchmark using a few assets*, Oper. Res., 54, pp. 232–246, <https://doi.org/10.1287/opre.1050.0260>.
- H. Y. ZHAO, W. P. TANG, AND D. YAO (2023), *Policy optimization for continuous reinforcement learning*, in 37th Advances in Neural Information Processing Systems (NeurIPS), Curran Associates, Red Hook, NY, pp. 1–27.