Astin Bulletin (2025), pp. 1–27
doi:10.1017/asb.2025.10080

RESEARCH ARTICLE

A nonzero-sum game with reinforcement learning under mean-variance framework

Junyi Guo[1], Xia Han[2], Hao Wang[3] and Kam C. Yuen[4]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, China
[2]School of Mathematical Sciences, LPMC and AAIS, Nankai University, Tianjin, 300071, China
[3]School of Mathematical Sciences, Nankai University, Tianjin, 300071, China
[4]Department of Statistics and Actuarial Science, The University of Hong Kong, Pok Fu Lam, Hong Kong
Corresponding author: Hao Wang; Email: hao.wang@mail.nankai.edu.cn

Received: 4 December 2024; **Revised:** 5 November 2025; **Accepted:** 6 November 2025

Keywords: Stochastic processes; reinforcement learning; Choquet regularizers; mean-variance framework; time-consistent Nash equilibrium

Abstract

In this paper, we investigate a competitive market involving two agents who consider both their own wealth and the wealth gap with their opponent. Both agents can invest in a financial market consisting of a risk-free asset and a risky asset, under conditions where model parameters are partially or completely unknown. This setup gives rise to a nonzero-sum differential game within the framework of reinforcement learning (RL). Each agent aims to maximize his own Choquet-regularized, time-inconsistent mean-variance objective. Adopting the dynamic programming approach, we derive a time-consistent Nash equilibrium strategy in a general incomplete market setting. Under the additional assumption of a Gaussian mean return model, we obtain an explicit analytical solution, which facilitates the development of a practical RL algorithm. Notably, the proposed algorithm achieves uniform convergence, even though the conventional policy improvement theorem does not apply to the equilibrium policy. Numerical experiments demonstrate the robustness and effectiveness of the algorithm, underscoring its potential for practical implementation.

1. Introduction

Since the seminal work of Markowitz (1952), the mean-variance criterion has emerged as a cornerstone in mathematical finance, highlighted by significant contributions such as Li and Ng (2000) and Zhou and Li (2000). It is well known that mean-variance problem has an inherent issue of time inconsistency, thereby Bellman optimality principle cannot be applied. Many studies addressing this issue resort to a pre-commitment strategy at the initial time and employ the Lagrangian method to solve the mean-variance problem as seen in the aforementioned works. However, the pre-committed solution lacks time consistency and is applicable only for a decision-maker at the initial time.

In pursuit of a time-consistent strategy, Björk and Murgoci (2010) formulated the problem within a game theoretic framework and derived the extended Hamilton–Jacobi–Bellman (HJB) equation through a verification theorem. Based on this groundwork, Björk *et al*. (2014) assumed that the insurer's risk aversion is inversely proportional to the current wealth and obtained the time-consistent strategy. Björk *et al*. (2017) provided a general framework for handling time-inconsistency, leading to the so-called equilibrium policy that can be regarded as a subgame perfect Nash equilibrium in dynamic games. Expanding on these efforts, Dai *et al*. (2021) proposed a dynamic portfolio choice model with the mean-variance criterion for log-returns, and derived time-consistent portfolio policies which are analytically tractable even under some incomplete market settings. Furthermore, the game theoretic approach to

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

tackling time-inconsistent problems has also been extended in the actuarial literature, as demonstrated by works from Zeng *et al*. (2016), Chen and Shen (2019), and Li and Young (2021).

In this paper, we consider the scenario where model parameters are partially or completely unknown, aiming to learn a practical exploratory equilibrium policy under the mean-variance criterion within an incomplete market, as discussed in Dai *et al*. (2023). In fact, the stochastic control problems under the continuous-time RL framework with continuous state and action have attracted extensive attention from scholars recently. Wang *et al*. (2020a) first established a continuous-time RL framework with continuous state and action from the perspective of stochastic control and proved that the optimal exploration strategy for the linear-quadratic (LQ) control problem in the infinite time horizon is Gaussian. Furthermore, Wang and Zhou (2020) applied this RL framework for the first time to solve the continuous-time mean-variance problem. Motivated by Wang *et al*. (2020a), Dai *et al*. (2023) extended the exploratory stochastic control framework to an incomplete market, where the asset return correlates with a stochastic market state, and learned an equilibrium policy under a mean-variance criterion. Jiang *et al*. (2022) studied the exploratory Kelly problem by considering both the amount of investment in stock and the portion of wealth in stock as the control for a general time-varying temperature parameter. Han *et al*. (2023) first introduced another kind of index that can measure the randomness of actions called Choquet regularization. They showed that the optimal exploration distribution of LQ control problem with infinite time horizon is no longer necessarily Gaussian as in Wang *et al*. (2020a), but are dictated by the choice of Choquet regularizers. Guo *et al*. (2025) further studied an exploratory mean-variance problem with the Choquet regularizers being used to measure the level of exploration.

Interactions among agents in real-world settings frequently involve a complex interplay of both competitive and cooperative behaviors. In contrast to the frameworks considered in the aforementioned literature, this paper focuses on a scenario with two competitive agents who evaluate not only their individual wealth but also the wealth gap relative to their opponent. This formulation aligns with the broader class of nonzero-sum stochastic differential games, tracing its foundational roots to seminal contributions by Isaacs (1965) and Pontryagin (1967).

In financial markets, investors often exhibit concern for their wealth relative to that of other market participants – a phenomenon commonly referred to as relative performance concern. This behavioral trait significantly influences investment decisions and market dynamics. Empirical and theoretical studies have documented that relative performance concerns can exacerbate speculative behavior and contribute to the emergence and persistence of asset bubbles. For example, DemMarzo *et al*. (2008) demonstrated that investors' concern with their relative performance induces herd behavior, thereby sustaining bubbles even when asset prices deviate significantly from their fundamental values. Foundational analyses by Abel (1990) and Gali (1994) emphasized the pervasive role of relative concerns in human decision-making processes. Within the portfolio optimization literature, Espinosa and Touzi (2015) incorporated concerns about relative wealth by modeling the wealth gap and showed that greater emphasis on relative wealth often leads investors to take on more risk, thus increasing systemic market risk. Building on this framework, numerous subsequent studies explicitly model objective functions that integrate weighted terms of individual wealth and wealth gaps, including Bensoussan *et al*. (2014), Siu *et al*. (2016), Hu and Wang (2018), Deng *et al*. (2018), Zhu *et al*. (2020), and Wang *et al*. (2021). To the best of our knowledge, (non)zero-sum games in the continuous-time RL settings have not been considered before except in Sun and Jia (2023) where an entropy-regularized continuous-time LQ two-agent zero-sum stochastic differential game problem was considered, and they designed a policy iteration method to derive the optimal strategy for a case with only one unknown model parameter.

Compared with the existing literature, this paper presents three main differences and contributions.

First, in traditional time-consistent optimization problems, policy iteration typically relies on the policy improvement theorem, as detailed in Jia and Zhou (2022b) and Guo *et al*. (2025). This theorem ensures that each iteration enhances the overall strategy. However, for time-inconsistent problems, although iterating policies is still feasible, there is no guarantee that each iteration leads to an improved policy. This presents a significant challenge when attempting to extend policy iteration methods to

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

time-inconsistent settings. Nevertheless, we show that, our approach guarantees uniform convergence of the equilibrium strategy, in contrast to the local convergence typically observed in the single-agent framework, as shown in Dai *et al*. (2023).

Second, unlike single-agent scenarios, our nonzero-sum game problem naturally fits within the domain of applying RL techniques in multi-agent systems. The pioneering work of Littman (1994) introduces Q-learning in zero-sum games, marking the early development of multi-agent reinforcement learning. Building on this, Littman (2001) proposed the Friend-or-foe Q-learning algorithm for general-sum games, while Foerster *et al*. (2017) proposed a centralized multi-agent learning method using deep learning, which enhances the collaborative to improve agent collaboration in complex tasks. Although centralized algorithms theoretically guarantee convergence and stability, they often encounter practical challenges, including dimensionality explosion and increased system complexity. For further discussion on multi-agent algorithms, we refer to Yang and Wang (2020) and Zhang *et al*. (2021). In contrast, the unique structure of our model allows the differential stochastic game to be decomposed into two independent single-agent problems within a centralized multi-agent framework, thus mitigating the dimensionality explosion inherent in centralized methods. To the best of our knowledge, this paper is the first to address equilibrium policies in time-inconsistent problems within the context of reinforcement learning for nonzero-sum differential games.

Third, in contrast to Jiang *et al*. (2022), Dai *et al*. (2023), and Sun and Jia (2023), we replace the differential entropy used for regularization with Choquet regularizers. As noted in Han *et al*. (2023) and Guo *et al*. (2025), Choquet regularizers offer several theoretical and practical advantages to RL. The broad class of Choquet regularizers enable the comparison and selection of specific regularizers to meet the unique objectives of each learning problem. In particular, it is more natural for agents to choose different regularizers based on their individual preferences, further enhancing the flexibility and applicability of our approach.

The rest of the paper is organized as follows. In Section 2, we introduce the exploratory mean-variance problem within the framework of RL under the nonzero-sum differential game setting. Section 3 presents the Nash equilibrium mean-variance policy. In Section 4, we show a policy iteration procedure and analyze its convergence based on the Gaussian mean return model. In Section 5, we propose an RL algorithm based on the convergence analysis, and provide numerical results to illustrate the implementation of the algorithm in Section 6. Finally, we conclude in Section 7. A summary of the notation used throughout the paper is provided in the Appendix.

2. Formulation of problem

Throughout the paper, we assume that $(\Omega, \mathscr{F}, \mathbb{P})$ be an atomless probability space. With a slight abuse of notation, we denote by $\mathscr{M}$ both the set of Borel probability measures on $\mathbb{R}$ and the set of distribution functions of real random variables. For $\Pi \in \mathscr{M}$ and $x \in \mathbb{R}$, we have $\Pi(x) = \Pi((-\infty, x])$. Let $\mathscr{M}^p$, $p \in [1, \infty)$, be the subset of $\mathscr{M}$ whose elements have finite p-th order moment. We write $X \sim \Pi$ if random variable X has distribution Π, and $X \overset{d}{=} Y$ if X and Y have the same distribution. For $\Pi \in \mathscr{M}^2$, we denote by $\mu(\Pi)$ and $\sigma^2(\Pi)$ the mean and variance of Π, respectively.

2.1 Exploratory wealth process

In the financial market under study, we posit the inherent incompleteness of the market as in Basak and Chabakauri (2010) and Dai *et al*. (2021), allowing for continuous trading in both a risk-free asset and a risky asset within the finite time horizon $[0, T]$. The price process $S_0(t)$ governing the risk-free asset is given by

$$\mathrm{d}S_0(t) = rS_0(t)\mathrm{d}t,$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

where $r > 0$ represents the constant risk-free interest rate. The price process $S(t)$ of the risky asset is given by

$$\frac{\mathrm{d}S(t)}{S(t)} = a(t, Y(t))\mathrm{d}t + b(t, Y(t))\mathrm{d}W(t). \tag{2.1}$$

Here, $W(t)$ is a Brownian motion defined on the filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{0\leqslant t\leqslant T}, \mathbb{P})$ adhering to standard conditions. Additionally, $Y(t)$ is a diffusion process satisfying

$$\mathrm{d}Y(t) = m(t, Y(t))\mathrm{d}t + v(t, Y(t))[\rho dW(t) + \sqrt{1-\rho^2}d\widetilde{W}(t)],$$

where $\rho \in [-1, 1]$, and the Brownian motion $\widetilde{W}(t)$ is defined on $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{0\leqslant t\leqslant T}, \mathbb{P})$ independent of $W(t)$. In the market context, $a(\cdot, \cdot)$, $b(\cdot, \cdot)$, $m(\cdot, \cdot)$, and $v(\cdot, \cdot)$ are deterministic functions of t and y, with all randomness entering through $Y(t)$. However, these functions themselves are unknown. Particularly, we are in the case of the complete market with constant parameters, where $a(t, y) \equiv a$ and $b(t, y) \equiv b$. The process $Y(t)$ is modeled as a diffusion representing macroeconomic or systemic risk factors. The parameter ρ quantifies the correlation between the state process $Y(t)$ and the stock price process $S(t)$, thus characterizing the degree to which market-wide risk factors influence asset price dynamics.

Remark 1. Market incompleteness arises because the stochastic evolution of Y(t) affects the dynamics of the risky asset, yet cannot be fully hedged through trading in S(t) and $S_0(t)$. This feature aligns with empirical observations, where financial markets exhibit unspanned stochastic volatility, unhedgeable economic risks, and other uncertainties. The market model described above exhibits substantial generality, encompassing numerous well-known models, such as the time-varying Gaussian mean return model and the stochastic volatility model; see Example 1 below. For a more detailed discussion on market incompleteness and the nonreplicability of contingent claims in such settings, we refer to Sections 1.4 and 5.4 of Föllmer and Schied (2011).

Example 1.

(i) *If the stock price S(t) and the market price of risk Y(t) are governed by* $\frac{\mathrm{d}S(t)}{S(t)} = (r + \sigma Y(t))\mathrm{d}t + \sigma \mathrm{d}W(t)$ *and* $\mathrm{d}Y(t) = \iota\,(Y - Y(t))\,\mathrm{d}t + v\mathrm{d}[\rho dW(t) + \sqrt{1-\rho^2}d\widetilde{W}(t)]$, *where* $r, \sigma, \iota, v,\ Y$ *are all positive constants and* $\rho \in [-1, 1]$. *This model, termed the time-varying Gaussian mean return model, effectively captures the intricate interplay between the dynamics of stock prices and market risk over time.*

(ii) *If the stock price S(t) and a state variable Y(t) follow* $\frac{\mathrm{d}S(t)}{S(t)} = (r + \sigma Y(t)^{\frac{1+\alpha}{2\alpha}})\mathrm{d}t + Y(t)^{\frac{1}{2\alpha}}\mathrm{d}W(t)$ *and* $\mathrm{d}Y(t) = \iota(Y - Y(t))dt + v\sqrt{Y(t)}\mathrm{d}[\rho dW(t) + \sqrt{1-\rho^2}d\widetilde{W}(t)]$, *where* $\alpha \neq 0$ *is the constant elasticity of the market price of risk* $\sigma Y(t)^{\frac{1}{2}}, \sigma \in \mathbb{R}, \iota > 0, v > 0,\ Y \in \mathbb{R}$ *and* $\rho \in [-1, 1]$ *are all constants, then it is a stochastic volatility model. This formulation characterizes a stochastic volatility model, acknowledging the nonlinear relationship between the state variable and market risk elasticity.*

Researchers, including Merton (1980), Kim and Omberg (1996), Liu (2001), Wachter (2002), Basak and Chabakauri (2010), Dai *et al.* (2021), and Dai *et al.* (2023), have extensively explored dynamic portfolio choice problems within the framework of these two market settings and their specific variations.

In what follows, we consider a nonzero-sum game involving two competing agents or companies, referred as Agent 1 and Agent 2 for simplicity. Both agents have access to the financial market. For $i \in \{1, 2\}$, let $u_i(t)$ represent the discounted amount of Agent i invested in the risky asset at time t, and the rest of the wealth is invested in the risk-free asset. Define $\boldsymbol{u}_i = \{u_i(t), 0 \leqslant t \leqslant T\}$ and

$$\theta(t, Y(t)) = \frac{a(t, Y(t)) - r}{b(t, Y(t))}. \tag{2.2}$$

Then the dynamic of discounted wealth process of Agent i under strategy u_i is given as

$$\mathrm{d}X_i^{u_i}(t) = u_i(t)b(t, Y(t))[\theta(t, Y(t))\mathrm{d}t + \mathrm{d}W(t)]. \tag{2.3}$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

Based on Wang *et al*. (2020a), we extend the control process (2.3) to a distributional control process and define the exploratory discounted wealth process for agents. Denote by $\Pi_i(t) \in \mathscr{M}^2$ the probability distribution function of the control u_i at time t, where $\mathscr{M}^2$ represents the set of distribution functions with finite second-order moment on $\mathbb{R}$. Let $\boldsymbol{\Pi}_i = \{\Pi_i(t), 0 \leqslant t \leqslant T\}$ and we write $\boldsymbol{\Pi}_i \in \mathscr{M}^2$ for simplicity. In contrast to the approach presented in Wang *et al*. (2020a), we additionally consider the correlation between the discounted wealth process and the market state, as well as the correlation between the two agents.

Now, we attempt to derive the exploratory version of the wealth process $X_i^{u_i}$ associated with randomized policy Π_i. Such a structure can also be found in Dai *et al*. (2023), Wang *et al*. (2020a), Wang and Zhou (2020). Let $W^n(t)$ and $\widetilde{W}^n(t), n = 1, 2, ..., N$, represent N paths independently sampled from $W(t)$ and $\widetilde{W}(t)$, respectively. Moreover, let $X_i^n(t)$ be the copies of the discounted wealth process of Agent i under strategy u_i^n sampled from Π_i. Then, for $n = 1, 2, ..., N$ and $i \in \{1, 2\}$, the increments of $Y^n(t)$ and the corresponding $X_i^n(t)$ can be written as

$$\begin{aligned}\Delta Y^n(t) &\equiv Y^n(t + \Delta t) - Y^n(t)\\ &\approx m(t, Y^n(t))\Delta t + v(t, Y^n(t))[\rho \Delta W^n(t) + \sqrt{1 - \rho^2}\Delta \widetilde{W}^n(t)],\end{aligned} \tag{2.4}$$

and

$$\begin{aligned}\Delta X_i^n(t) &\equiv X_i^n(t + \Delta t) - X_i^n(t)\\ &\approx u_i^n(t)b(t, Y^n(t))[\theta(t, Y^n(t))\Delta t + \Delta W^n(t)].\end{aligned} \tag{2.5}$$

We denote the exploratory discounted wealth process of Agent i by $X_i^{\Pi_i}(t)$. Consequently, $X_i^n(t)$ can be viewed as an independent sample from $X_i^{\Pi_i}(t)$. By the law of large numbers and using (2.4) and (2.5), we observe that, as $N \to \infty$,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \Delta X_i^n(t) &\approx \frac{1}{N}\sum_{n=1}^{N} u_i^n(t)b(t, Y^n(t))[\theta(t, Y^n(t))\Delta t + \Delta W^n(t)]\\ &\xrightarrow{a.s.} \mathbb{E}\left[b(t, Y(t))\theta(t, Y(t)) \int_U u \mathrm{d}\Pi_i(t, u)\right] \Delta t,\end{aligned} \tag{2.6}$$

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} (\Delta X_i^n(t))^2 &\approx \frac{1}{N}\sum_{n=1}^{N} (u_i^n(t)b(t, Y^n(t)))^2 \Delta t\\ &\xrightarrow{a.s.} \mathbb{E}\left[b^2(t, Y(t)) \int_U u^2 \mathrm{d}\Pi_i(t, u)\right] \Delta t,\end{aligned} \tag{2.7}$$

and

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \Delta X_i^n(t)\Delta Y^n(t) &\approx \frac{1}{N}\sum_{n=1}^{N} u_i^n(t)b(t, Y^n(t))\rho v(t, Y^n(t))\Delta t\\ &\xrightarrow{a.s.} \mathbb{E}\left[\rho b(t, Y(t))v(t, Y(t)) \int_U u \mathrm{d}\Pi_i(t, u)\right] \Delta t.\end{aligned} \tag{2.8}$$

It is well known from the law of large numbers that

$$\frac{1}{N}\sum_{n=1}^{N} \Delta X_i^n(t) \xrightarrow{a.s.} \mathbb{E}[\Delta X_i^{\Pi_i}(t)],\ \frac{1}{N}\sum_{n=1}^{N} (\Delta X_i^n(t))^2 \xrightarrow{a.s.} \mathbb{E}[(\Delta X_i^{\Pi_i}(t))^2],$$

and

$$\frac{1}{N}\sum_{n=1}^{N} \Delta X_i^n(t)\Delta Y^n(t) \xrightarrow{a.s.} \mathbb{E}[\Delta X_i^{\Pi_i}(t)\Delta Y(t)].$$

These together with (2.6)–(2.8), motivate our confidence in the dynamic of $\Delta X_i^{\Pi_i}(t)$, which can be expressed as

$$\mathrm{d}X_i^{\Pi_i}(t) = b(t, Y(t))\theta(t, Y(t))\mu_i(t)\mathrm{d}t + b(t, Y(t))[\mu_i(t)\mathrm{d}W(t) + \sigma_i(t)\mathrm{d}\overline{W}_i(t)], \tag{2.9}$$

where

$$\mu_i(t) = \int_U u\mathrm{d}\Pi_i(t, u), \;\; \sigma_i^2(t) = \int_U u^2\mathrm{d}\Pi_i(t, u) - \mu_i^2(t)$$

with $\overline{W}_i$ being the Brownian motion independent of $W(t)$ and $\widetilde{W}(t)$. The remaining consideration involves the correlation between $\overline{W}_1(t)$ and $\overline{W}_2(t)$. It is observed that, as $N \to \infty$,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\Delta X_1^n(t)\Delta X_2^n(t) &\approx \frac{1}{N}\sum_{n=1}^{N} u_1^n(t)u_2^n(t)b^2(t, Y^n(t)) \\ &\xrightarrow{a.s.} \mathbb{E}\left[\mu_1(t)\mu_2(t)b^2(t, Y(t))\right]\Delta t,\end{aligned}$$

leading to $\langle \overline{W}_1, \overline{W}_2\rangle = 0$. By Lévy's theorem, we conclude that $\overline{W}_1$ is independent of $\overline{W}_2$.

Remark 2. Indeed, the above construction follows from the framework of relaxed stochastic control. In classical control theory, the strategy $u(\omega, t)$ is a stochastic process that takes a deterministic value for fixed ω and t. When we randomize the strategy $u(\omega, t)$ to $\Pi(\omega, t)$, then for each fixed ω and t, $\Pi(\omega, t)$ is a distribution on U, which we write as $\Pi_t(\omega, \mathrm{d}u)$. For any t, we can view $\Pi_t(\omega, B)$ with $B \in \mathscr{B}(U)$, as a probability kernel with source $(\Omega, \mathscr{F}_t)$ and target $(U, \mathscr{B}(U))$. By standard measure-theoretic results, $\Pi_t(\omega, \cdot)$ together with the probability $\mathbb{P}$ on $(\Omega, \mathscr{F}_t)$ uniquely induces a measure $\mathbb{Q}$ on $(\Omega \times U, \mathscr{F}_t \times \mathscr{B}(U))$ satisfying $\mathbb{Q}(\mathrm{d}\omega, \mathrm{d}u) = \Pi_t(\omega, \mathrm{d}u)\mathbb{P}(\mathrm{d}\omega)$. The sampling described above is thus carried out on the probability space $(\Omega \times U, \mathscr{F}_t \times \mathscr{B}(U), \mathbb{Q})$. Therefore, by the generalized Fubini theorem, for any measurable function f on $(\Omega \times U, \mathscr{F}_t \times \mathscr{B}(U), \mathbb{Q})$, we have

$$\mathbb{E}^{\mathbb{Q}}[f] = \int_{\Omega\times U} f\mathrm{d}\mathbb{Q} = \int_\Omega\int_U f(\omega, u)\Pi_t(\omega, \mathrm{d}u)\mathrm{d}\mathbb{P} = \mathbb{E}\left[\int_U f(\omega, u)\Pi_t(\omega, \mathrm{d}u)\right].$$

For $i, j \in \{1, 2\}$ and $j \neq i$, assume that Agent i takes into account not only his own wealth but also the wealth gap between himself and Agent j at the terminal time T. Given a strategy $\Pi_j \in \mathscr{M}^2$ employed by Agent j, Agent i will choose a strategy Π_i to maximize the following objective

$$\begin{aligned}&\mathbb{E}_t[(1-k_i)X_i^{\Pi_i}(T) + k_i(X_i^{\Pi_i}(T) - X_j^{\Pi_j}(T))] - \frac{\gamma_i}{2}\mathrm{Var}_t[(1-k_i)X_i^{\Pi_i}(T) + k_i(X_i^{\Pi_i}(T) - X_j^{\Pi_j}(T))] \\ &= \mathbb{E}_t[X_i^{\Pi_i}(T) - k_iX_j^{\Pi_j}(T)] - \frac{\gamma_i}{2}\mathrm{Var}_t[X_i^{\Pi_i}(T) - k_iX_j^{\Pi_j}(T)],\end{aligned} \tag{2.10}$$

where $\mathbb{E}_t(\cdot)$ and $\mathrm{Var}_t(\cdot)$ denote the conditional expectation and variance with given $X_1^{\Pi_1}(t)$, $X_2^{\Pi_2}(t)$ and $Y(t)$, respectively, $\gamma_i > 0$ represents the risk-aversion coefficient for Agent i, and $k_i \in (0, 1)$ measures the sensitivity of Agent i to the performance of Agent j. Notably, a larger k_i implies that Agent i emphasis on the relative performance against his opponent (Agent j), thereby intensifying the competitiveness of the game. Similar models addressing relative performance concerns, which do not incorporate RL, have been studied in the literature. For related discussions, see, for example, Bensoussan *et al.* (2014), as well as the relevant references cited in our introduction.

Remark 3. Unlike the objective of maximizing the insurer's utility in a nonzero-sum game (see, e.g., Browne, 2000 and Bensoussan *et al.* 2014), we consider a mean-variance objective as in Wang *et al.* (2019) and (2021). On one hand, the agents seek to maximize the expected wealth gap; on the other hand, they prioritize stability and predictability over excessive volatility. By minimizing the variance of the wealth gap, the agent reduces uncertainty in their relative wealth position, aligning with standard assumptions of risk aversion.

Let $\hat{X}_i^{\Pi_i,\Pi_j}(t) = X_i^{\Pi_i}(t) - k_i X_j^{\Pi_j}(t)$ be the wealth difference of the two agents. It is obvious from (2.9) that $\hat{X}_i^{\Pi_i,\Pi_j}(t)$ follows the dynamic

$$\begin{aligned} \mathrm{d}\hat{X}_i^{\Pi_i,\Pi_j}(t) = b(t, Y(t))[&\theta(t, Y(t))(\mu_i(t) - k_i\mu_j(t))\mathrm{d}t \\ &+ (\mu_i(t) - k_i\mu_j(t))\mathrm{d}W(t) + \sigma_i(t)\mathrm{d}\overline{W}_i(t) - k_i\sigma_j(t)\mathrm{d}\overline{W}_j(t)]. \end{aligned} \tag{2.11}$$

2.2 Objective function

We employ the *Choquet regularizer* Φ_h to quantify randomness. Given a concave function $h : [0, 1] \to \mathbb{R}$ of bounded variation with $h(0) = h(1) = 0$ and $\Pi \in \mathscr{M}$, the Choquet regularizer Φ_h on $\mathscr{M}$ is defined as

$$\Phi_h(\Pi) = \int_{\mathbb{R}} h \circ \Pi([x, \infty))\mathrm{d}x. \tag{2.12}$$

We denote the set of $h : [0, 1] \to \mathbb{R}$ by $\mathscr{H}$. In fact, (2.12) is a signed Choquet integral characterized by Wang *et al*. (2020c) via comonotonic additivity, which essentially builds on the seminal works of Schmeidler (1989) and Yaari (1987).

As stated in Lemma 2.2 of Han *et al*. (2023), the regularizer Φ_h is rigorously defined and serves as an important metric for quantifying the degree of randomness or exploration within the context of RL. Specifically, the concavity of h ensures that Φ_h is also concave. This means that $\Phi_h(\lambda\Pi_1 + (1-\lambda)\Pi_2) \geqslant \lambda\Phi_h(\Pi_1) + (1-\lambda)\Phi_h(\Pi_2)$ for all $\Pi_1, \Pi_2 \in \mathscr{M}$ and $\lambda \in [0, 1]$, which intuitively implies that the linear combination of two distributions is more random. Moreover, the condition $h(0) = h(1) = 0$ means for any $c \in \mathbb{R}$, $\Phi_h(\delta_c) = 0$, where δ_c is the Dirac mass at c. This indicates that degenerate distributions do not have any randomness measured by Φ_h. Additionally, the agents have the flexibility to opt for different regularizers, contingent upon their preferences, as reflected by the distortion function h. For more detailed discussions about the properties associated with Φ_h, we refer to Han *et al*. (2023) and Guo *et al*. (2025).

It is useful to note that Φ_h admits a quantile representation, see Lemma 1 of Wang *et al*. (2020b). For a distribution $\Pi \in \mathscr{M}$, let its left-quantile for $p \in (0, 1]$ be defined as

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

then we have

$$\Phi_h(\Pi) = \int_0^1 Q_\Pi(1-p)\mathrm{d}h(p) \tag{2.13}$$

if h is left-continuous.

For any fixed $\Pi_j \in \mathscr{M}^2$, we incorporate the Choquet regularizer Φ_{h_i} for Agent i, along with the exploration weight function $\lambda_i(t)$, into the mean-variance criterion (2.10). This regularizer plays a role analogous to differential entropy in entropy-regularized reinforcement learning (e.g., Wang *et al*., 2020a; Wang and Zhou, 2020), serving to prevent the learned policy from collapsing to a deterministic solution. Thus, each agent aims to achieve an exploratory mean-variance problem within the framework of RL. This yields the corresponding objective function

$$J_i(t, \hat{x}_i, y; \mathbf{\Pi}_i, \mathbf{\Pi}_j) := \mathbb{E}_t\left[\int_t^T \lambda_i(s)\Phi_{h_i}(\Pi_i(s))ds + \hat{X}_i^{\Pi_i,\Pi_j}(T)\right] - \frac{\gamma_i}{2}\mathrm{Var}_t[\hat{X}_i^{\Pi_i,\Pi_j}(T)], \tag{2.14}$$

where $\hat{X}_i^{\Pi_i,\Pi_j}(t) = \hat{x}_i$ and $Y(t) = y$ with t representing the initial time.

Remark 4. We note that a larger $\lambda_i(t)$ promotes increased exploration, as it results in a higher weight on $\Phi_{h_i}(\Pi_i(t))$. When $\lambda_i(t) \equiv \lambda_i$, the exploration weight remains constant over time. It is often more realistic to set $\lambda_i(t)$ to decrease over time; for instance, $\lambda(t)$ may follow a power-decaying pattern, expressed as $\lambda_i(t) = \lambda_0(T+\lambda)^{\lambda_0}/(t+\lambda)^{\lambda_0+1}$ with $\lambda_0, \lambda > 0$. Alternatively, $\lambda_i(t)$ may decay exponentially with $\lambda_i(t) = \lambda_0 e^{\lambda_0(T-t)}$ with $\lambda_0 > 0$. For further discussions on selecting $\lambda_i(t)$, we refer to Section 3.4 of Jiang *et al.* (2022).

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

For $i, j \in \{1, 2\}$ and $i \neq j$, denote by $\mathscr{A}(\mathbf{\Pi}_j)$ the set of all admissible feedback control of $\mathbf{\Pi}_i$. Given $\mathbf{\Pi}_j \in \mathscr{M}^2$, $\mathbf{\Pi}_i$ is said to be in $\mathscr{A}(\mathbf{\Pi}_j)$ if the following conditions hold:

(i) For each $s \in [t, T]$, $\Pi_i(s) \in \mathscr{M}^2(\mathbb{R})$;

(ii) There exists a deterministic mapping $\xi_i : [t, T] \times \mathbb{R} \times \mathbb{R} \to \mathscr{M}^2(\mathbb{R})$, such that $\Pi_i(s) = \xi_i(s, \hat{X}_i^{\Pi_i,\Pi_j}(s), Y(s))$;

(iii) For any $A \in \mathscr{B}(\mathbb{R})$, $\{\int_A \Pi_i(s, u)\mathrm{d}u, t \leqslant s \leqslant T\}$ is $\mathscr{F}_s$–progressively measurable;

(iv) $\mathbb{E}_t \int_t^T [|b(s, Y(s))\theta(s, Y(s))(\mu_i(s) - k_i\mu_j(s))| + b^2(s, Y(s))(\mu_i^2(s) + \sigma_i^2(s))]ds < \infty$;

(v) $\mathbb{E}_t \int_t^T |\lambda_i(s)\Phi_{h_i}(\Pi_i(s))|ds < \infty$.

Next, we define the *profile* of the game, which encapsulates the comprehensive strategic behavior of both agents in the game.

Definition 1. *A strategy pair* $(\mathbf{\Pi}_1, \mathbf{\Pi}_2)$ *is called the profile of the game if it satisfies the following conditions: given the strategy* $\mathbf{\Pi}_2$, $\mathbf{\Pi}_1$ *is an admissible feedback control, and conversely, given the strategy* $\mathbf{\Pi}_1$, $\mathbf{\Pi}_2$ *is also an admissible feedback control.*

As mentioned in the introduction, to seek the time-consistent equilibrium strategy, we formulate the time-inconsistent dynamic optimization problem into a noncooperative game theoretic framework proposed by Björk and Murgoci (2014) and Björk *et al*. (2017).

Definition 2. *Let* $(\mathbf{\Pi}_1, \mathbf{\Pi}_2)$ *be a profile of the game. For* $i, j \in \{1, 2\}$ *and* $i \neq j$, $\mathbf{\Pi}_i$ *is said to be equilibrium response of Agent i if for a fixed* Δ *and for any initial state* $(t, \hat{x}_i, y)$ *and an arbitrary* $\eta_i \in \mathscr{M}^2$, $\mathbf{\Pi}_i^{\eta_i,\Delta}$ *defined by*

$$\Pi_i^{\eta_i,\Delta}(s) = \begin{cases} \eta_i, & t \leqslant s \leqslant t + \Delta, \\ \Pi_i(s), & t + \Delta \leqslant s \leqslant T, \end{cases}$$

satisfying

$$\liminf_{\Delta \to 0^+} \frac{J_i(t, \hat{x}_i, y; \mathbf{\Pi}_i, \mathbf{\Pi}_j) - J_i(t, \hat{x}_i, y; \mathbf{\Pi}_i^{\eta_i,\Delta}, \mathbf{\Pi}_j)}{\Delta} \geqslant 0.$$

The equilibrium response $\mathbf{\Pi}_i$ *of Agent i can be viewed as a mapping of* $\mathbf{\Pi}_j$, *and thus can be written as* $\mathbf{\Pi}_i = \kappa_i(\mathbf{\Pi}_j)$ *at this point. Furthermore, the equilibrium response value function of Agent i is defined as*

$$\widetilde{V}_i(t, \hat{x}_i, y; \mathbf{\Pi}_j) := J_i(t, \hat{x}_i, y; \mathbf{\Pi}_i, \mathbf{\Pi}_j).$$

3. Time-consistent Nash equilibrium

In this section, we present the Nash equilibrium for a general incomplete market. It is important to emphasize that the uniqueness of the equilibrium policy remains an open question in the context of time-inconsistent optimization problems, as discussed by Ekeland and Pirvu (2008). Therefore, within the framework of game theory, we focus on a specific Nash equilibrium defined below.

Definition 3. *A profile* $(\mathbf{\Pi}_1^*, \mathbf{\Pi}_2^*)$ *is called the time-consistent Nash equilibrium of the game if for* $i, j \in \{1, 2\}$ *and* $i \neq j$, $\mathbf{\Pi}_i^*$ *is the equilibrium response of Agent i, that is* $\mathbf{\Pi}_1^* = \kappa_1(\mathbf{\Pi}_2^*)$ *and* $\mathbf{\Pi}_2^* = \kappa_2(\mathbf{\Pi}_1^*)$. *Furthermore, the equilibrium value function of Agent i is given by*

$$V_i(t, \hat{x}_i, y) := \widetilde{V}_i(t, \hat{x}_i, y; \mathbf{\Pi}_j^*) = J_i(t, \hat{x}_i, y; \mathbf{\Pi}_i^*, \mathbf{\Pi}_j^*).$$

3.1 Verification theorem

Analogous to the work of Björk *et al*. (2017) on optimization problems involving a broad class of objective functionals, we present the following verification theorem.

Let $\mathscr{D} = [0, T] \times \mathbb{R}^2$ and then for any $\varphi \in C^{1,2,2}(\mathscr{D})$, we can denote the infinitesimal generator of $(\hat{X}^{\Pi_i,\Pi_j}(t), Y(t))$ by

$$
\begin{aligned}
\mathscr{L}^{\Pi_i,\Pi_j}\varphi(t,x,y)=&\frac{\partial\varphi}{\partial t}+b(t,y)\theta(t,y)(\mu_i(t)-k_i\mu_j(t))\frac{\partial\varphi}{\partial x}+\frac{1}{2}b^2(t,y)\left((\mu_i(t)-k_i\mu_j(t))^2\right.\\
&\left.+\sigma_i^2(t)+k_i^2\sigma_j^2(t)\right)\frac{\partial^2\varphi}{\partial x^2}+m(t,y)\frac{\partial\varphi}{\partial y}+\frac{1}{2}v^2(t,y)\frac{\partial^2\varphi}{\partial y^2}\\
&+\rho v(t,y)b(t,y)(\mu_i(t)-k_i\mu_j(t))\frac{\partial^2\varphi}{\partial x\partial y}.
\end{aligned}
\tag{3.1}
$$

Let $\hat{X}_i^{u_i,u_j}(t)=X_i^{u_i}(t)-k_iX_j^{u_j}(t)$ and denote the infinitesimal generator of $(\hat{X}_i^{u_i,u_j}(t),Y(t))$ by $\mathscr{L}^{u_i,u_j}$. It can be directly verified that $\mathscr{L}^{\Pi_i,\Pi_j}\varphi(t,x,y)=\int_{\mathbb{R}}\int_{\mathbb{R}}\mathscr{L}^{u_i,u_j}\varphi(t,x,y)\mathrm{d}\Pi_i(u_i)\mathrm{d}\Pi_j(u_j)$.

Theorem 1 (Verification theorem). *For $i,j\in\{1,2\}$ and $i\neq j$, fix $\mathbf{\Pi}_j$ and suppose that functions $V_i(t,x,y)\in C^{1,2,2}(\mathscr{D})$, $g_i(t,x,y)\in C^{1,2,2}(\mathscr{D})$ and strategy $\mathbf{\Pi}_i$ satisfy the following properties:*

(i) *V_i and g_i solve the extended HJB system*

$$
\sup_{\Pi_i\in\mathscr{M}^2}\left\{\mathscr{L}^{\Pi_i,\Pi_j}V_i(t,\hat{x}_i,y)-\frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j}g_i^2(t,\hat{x}_i,y)+\gamma_ig_i\mathscr{L}^{\Pi_i,\Pi_j}g_i(t,\hat{x}_i,y)+\lambda_i(t)\Phi_{h_i}(\Pi_i)\right\}=0,
\tag{3.2}
$$

and

$$
\mathscr{L}^{\Pi_i^*,\Pi_j}g_i(t,\hat{x}_i,y)=0
\tag{3.3}
$$

with

$$
V_i(T,\hat{x}_i,y)=\hat{x}_i,\;\; g_i(T,\hat{x}_i,y)=\hat{x}_i,
\tag{3.4}
$$

where $\mathscr{L}^{\Pi_i,\Pi_j}$ is the infinitesimal generator given by (3.1).

(ii) *Π_i^* realizes the supremum in (3.2) and $\mathbf{\Pi}_i^*=\{\Pi_i^*(t),0\leqslant t\leqslant T\}$ is admissible.*

Then $\mathbf{\Pi}_i^*$ is the equilibrium response of Agent i. Furthermore, $V_i(t,\hat{x}_i,y)=J_i(t,\hat{x}_i,y;\,\mathbf{\Pi}_i^*,\,\mathbf{\Pi}_j)$ is the equilibrium response value function of Agent i and $g_i(t,\hat{x}_i,y)=\mathbb{E}_t[\hat{X}_i^{\Pi_i^*,\Pi_j}(T)]$.

3.2 Solution to the general case

While the preservation of the uniqueness of equilibrium response remains uncertain, the aforementioned theorem provides a constructive framework for identifying a specific Nash equilibrium. We first concentrate on the equilibrium response of Agent 1, and analogous results for Agent 2 can be obtained using the same method. It is assumed that the equilibrium value function V_1 and the function g_1 can be precisely expressed as

$$
V_1(t,x,y)=x+D_1(t,y),\quad g_1(t,x,y)=x+d_1(t,y).
\tag{3.5}
$$

It is obvious that $D_1(T,y)=d_1(T,y)=0$. Using (3.1) and substituting (3.5) into (3.2), we streamline (3.2) to

$$
\begin{aligned}
\sup_{\Pi_1\in\mathscr{M}^2}\Bigg\{&\frac{\partial D_1(t,y)}{\partial t}+b(t,y)\theta(t,y)(\mu_1(t)-k_1\mu_2(t))+m(t,y)\frac{\partial D_1(t,y)}{\partial y}\\
&-\frac{\gamma_1}{2}b^2(t,y)[(\mu_1^2(t)+\sigma_1^2(t))+k_1^2(\mu_2^2(t)+\sigma_2^2(t))-2k_1\mu_1(t)\mu_2(t)]\\
&-\frac{\gamma_1}{2}v^2(t,y)\left(\frac{\partial d_1(t,y)}{\partial y}\right)^2+\frac{1}{2}v^2(t,y)\frac{\partial^2D_1(t,y)}{\partial y^2}\\
&-\gamma_1\rho v(t,y)b(t,y)(\mu_1(t)-k_1\mu_2(t))\frac{\partial d_1(t,y)}{\partial y}+\lambda_1(t)\Phi_{h_1}(\Pi_1)\Bigg\}=0.
\end{aligned}
\tag{3.6}
$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

We can see from (3.6) that the supremum only depends on the mean and variance of Π_1 except $\Phi_{h_1}(\Pi_1)$. We proceed with our analysis relying on the crucial Lemma 1.

Lemma 1 (Theorem 3.1 of Liu *et al.* 2020). *If h is continuous and not constantly zero, then a maximizer* Π^* *to the optimization problem*

$$\max_{\Pi\in\mathscr{M}^2} \Phi_h(\Pi) \qquad \textit{subject to } \mu(\Pi)=m \textit{ and } \sigma^2(\Pi)=s^2 \tag{3.7}$$

has the following quantile function

$$Q_{\Pi^*}(p)=m+s\frac{h'(1-p)}{||h'||_2}, \quad \textit{a.e. } p\in(0,1), \tag{3.8}$$

and the maximum value of (3.7) *is* $\Phi_h(\Pi^*)=s||h'||_2$.

Let $\Lambda_1(\Pi_1)$ denote the expression enclosed in braces in (3.6). We then have the following proposition.

Proposition 1. *Let* $h_1\in\mathscr{H}$ *be a continuous function. For any strategy* $\mathbf{\Pi}_1=\{\Pi_1(t),0\leqslant t\leqslant T\}\in\mathscr{A}(\mathbf{\Pi}_2)$ *with mean process* $\{\mu_1(t)\}_{0\leqslant t\leqslant T}$ *and variance process* $\{\sigma_1(t)^2\}_{0\leqslant t\leqslant T}$, *there exists a strategy* $\mathbf{\Pi}_1^*=\{\Pi_1^*(t),0\leqslant t\leqslant T\}\in\mathscr{A}(\mathbf{\Pi}_2)$ *defined by*

$$Q_{\Pi_1^*(t)}(p)=\mu_1(t)+\sigma_1(t)\frac{h_1'(1-p)}{||h_1'||_2}, \quad \textit{a.e. } p\in(0,1),\ \ 0\leqslant t\leqslant T,$$

which shares the same mean and variance processes as $\mathbf{\Pi}_1$ *and satisfies* $\Lambda_1(\Pi_1(t))\leqslant\Lambda_1(\Pi_1^*(t))$.

Proof. From (3.6), $\Lambda_1(\Pi_1)$ depends on Π_1 only through $\mu_1(\Pi_1)$, $\sigma_1^2(\Pi_1)$ and $\Phi_{h_1}(\Pi_1)$. Since $\mathbf{\Pi}_1^*$ has the same mean and variance processes as $\mathbf{\Pi}_1$, the problem reduces to finding $\mathbf{\Pi}_1^*$ that maximizes

$$\max_{\Pi_1(t)\in\mathscr{M}} \Phi_{h_1}(\Pi_1(t)) \ \text{ subject to } \mu_1(\Pi_1(t))=\mu_1(t),\ \ \sigma_1(\Pi_1(t))^2=\sigma_1(t)^2 \ \text{ for all } t\in[0,T]. \tag{3.9}$$

By Lemma 1, the maximizer $\Pi_1^*(t)$ of this problem has quantile function

$$Q_{\Pi_1^*(t)}(p)=\mu_1(t)+\sigma_1(t)\frac{h_1'(1-p)}{\|h_1'\|_2},$$

and satisfies $\Phi_{h_1}(\Pi_1^*(t))=\sigma_1(t)\|h_1'\|_2$. This completes the proof. □

To find the equilibrium response of Agent 1, we need to solve problem (3.6). Let $\mu_1^*(t)$ and $\sigma_1^*(t)$ represent the mean and standard deviation of equilibrium response of Agent 1. As in the proof of Proposition 1, with given $\mu_1(t)$ and $\sigma_1(t)$, $\Lambda_1(\Pi_1(t))$ only depends on $\Phi_{h_1}(\Pi_1(t))$. Thus, we can rewrite (3.6) as

$$\max_{\Pi_1\in\mathscr{M}}\Lambda_1(\Pi_1)=\max_{m\in\mathbb{R},s>0}\ \max_{\substack{\Pi_1\in\mathscr{M}\\ \mu_1(\Pi_1)=m,\sigma_1(\Pi_1)^2=s^2}}\Lambda_1(\Pi_1).$$

The inner maximization problem is equivalent to (3.7) or (3.9), and its maximizer Π_1 satisfies $\Phi_{h_1}(\Pi_1)=s\|h_1'\|_2$. Substituting it back into $\Lambda(\Pi_1)$, we obtain

$$\begin{aligned}(\mu_1^*(t),\sigma_1^*(t))=\arg\max_{m\in\mathbb{R},s>0}&\{b(t,y)\theta(t,y)(m-k_1\mu_2(t))\\ &-\frac{\gamma_1}{2}b^2(t,y)[(m^2+s^2)+k_1^2(\mu_2^2(t)+\sigma_2^2(t))-2k_1m\mu_2(t)]\\ &-\gamma_1\rho v(t,y)b(t,y)(m-k_1\mu_2(t))\frac{\partial d_1(t,y)}{\partial y}+\lambda_1(t)s\|h_1'\|_2\Big\}.\end{aligned}$$

By the first-order condition, we deduce that

$$b(t,y)\theta(t,y)-\gamma_1b^2(t,y)\mu_1^*(t)+k_1\gamma_1\mu_2(t)b^2(t,y)-\gamma_1\rho v(t,y)b(t,y)\frac{\partial d_1(t,y)}{\partial y}=0,$$

and

$$-\gamma_1b^2(t,y)\sigma_1^*(t)+\lambda_1(t)\|h_1'\|_2=0.$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

So the mean and standard deviation of equilibrium response of Agent 1 are

$$\mu_1^*(t) = \frac{\theta(t,y)}{\gamma_1 b(t,y)} + k_1\mu_2(t) - \frac{\rho v(t,y)}{b(t,y)}\frac{\partial d_1(t,y)}{\partial y}, \tag{3.10}$$

and

$$\sigma_1^*(t) = \frac{\lambda_1(t)\|h_1'\|_2}{\gamma_1 b^2(t,y)}. \tag{3.11}$$

By substituting (3.10) and (3.11) back into (3.3) and (3.6), we then get that $d_1(t,y)$ and $D_1(t,y)$, respectively, satisfy

$$\frac{\partial d_1(t,y)}{\partial t} + (m(t,y) - \rho v(t,y)\theta(t,y))\frac{\partial d_1(t,y)}{\partial y} + \frac{1}{2}v^2(t,y)\frac{\partial^2 d_1(t,y)}{\partial y^2} + \frac{\theta^2(t,y)}{\gamma_1} = 0,$$

and

$$\begin{aligned}
&\frac{\partial D_1(t,y)}{\partial t} + m(t,y)\frac{\partial D_1(t,y)}{\partial y} + \frac{1}{2}v^2(t,y)\frac{\partial^2 D_1}{\partial y^2}(t,y)\\
&-\frac{\gamma_1}{2}(1-\rho^2)v^2(t,y)\left(\frac{\partial d_1}{\partial y}(t,y)\right)^2 - \rho v(t,y)\theta(t,y)\frac{\partial d_1}{\partial y}\\
&+\frac{\theta^2(t,y)}{2\gamma_1} - \frac{\gamma_1 k_1^2}{2}b^2(t,y)\sigma_2^2(t) + \frac{\lambda_1^2(t)\|h_1'\|_2^2}{2\gamma_1 b^2(t,y)} = 0.
\end{aligned}$$

Repeating the procedure outlined above can yield results for Agent 2. So if $(\mathbf{\Pi}_1^*, \mathbf{\Pi}_2^*)$ constitutes a Nash equilibrium, we have

$$\begin{cases}
\mu_1^*(t) - k_1\mu_2^*(t) = \dfrac{\theta(t,y)}{\gamma_1 b(t,y)} - \dfrac{\rho v(t,y)}{b(t,y)}\dfrac{\partial d_1(t,y)}{\partial y},\\[2ex]
\mu_2^*(t) - k_2\mu_1^*(t) = \dfrac{\theta(t,y)}{\gamma_2 b(t,y)} - \dfrac{\rho v(t,y)}{b(t,y)}\dfrac{\partial d_2(t,y)}{\partial y}.
\end{cases} \tag{3.12}$$

The solution to the system of equations can be directly calculated as

$$\begin{cases}
\mu_1^*(t) = \dfrac{1}{1-k_1k_2}\left[\dfrac{\theta(t,y)}{b(t,y)}\left(\dfrac{1}{\gamma_1} + \dfrac{k_1}{\gamma_2}\right) - \dfrac{\rho v(t,y)}{b(t,y)}\left(\dfrac{\partial d_1(t,y)}{\partial y} + \dfrac{k_1\partial d_2(t,y)}{\partial y}\right)\right],\\[2ex]
\mu_2^*(t) = \dfrac{1}{1-k_1k_2}\left[\dfrac{\theta(t,y)}{b(t,y)}\left(\dfrac{1}{\gamma_2} + \dfrac{k_2}{\gamma_1}\right) - \dfrac{\rho v(t,y)}{b(t,y)}\left(\dfrac{\partial d_2(t,y)}{\partial y} + \dfrac{k_2\partial d_1(t,y)}{\partial y}\right)\right].
\end{cases} \tag{3.13}$$

Summarizing the above, we get following theorem.

Theorem 2. *For $i,j \in \{1,2\}$ and $i \neq j$, let $d_i(t,y)$ and $D_i(t,y)$ be the solutions of*

$$\frac{\partial d_i(t,y)}{\partial t} + (m(t,y) - \rho v(t,y)\theta(t,y))\frac{\partial d_i(t,y)}{\partial y} + \frac{1}{2}v^2(t,y)\frac{\partial^2 d_i}{\partial y^2} + \frac{\theta^2(t,y)}{\gamma_i} = 0, \tag{3.14}$$

and

$$\begin{aligned}
&\frac{\partial D_i(t,y)}{\partial t} + m(t,y)\frac{\partial D_i(t,y)}{\partial y} + \frac{1}{2}v^2(t,y)\frac{\partial^2 D_i}{\partial y^2}(t,y)\\
&-\frac{\gamma_i}{2}(1-\rho^2)v^2(t,y)\left(\frac{\partial d_i}{\partial y}(t,y)\right)^2 - \rho v(t,y)\theta(t,y)\frac{\partial d_i}{\partial y}(t,y)\\
&+\frac{\theta^2(t,y)}{2\gamma_i} - \frac{\gamma_i k_i^2}{2}b^2(t,y)\sigma_j^2(t) + \frac{\lambda_i^2(t)\|h_i'\|_2^2}{2\gamma_i b^2(t,y)} = 0,
\end{aligned} \tag{3.15}$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

with terminal condition $D_i(T,y)=d_i(T,y)=0$. *Then* $(\boldsymbol{\Pi}_1^*,\boldsymbol{\Pi}_2^*)$ *with quantile functions*

$$Q_{\Pi_i^*(t)}(p)=\frac{1}{1-k_1k_2}\left[\frac{\theta(t,y)}{b(t,y)}\left(\frac{1}{\gamma_i}+\frac{k_i}{\gamma_j}\right)-\frac{\rho v(t,y)}{b(t,y)}\left(\frac{\partial d_i(t,y)}{\partial y}+\frac{k_i\partial d_j(t,y)}{\partial y}\right)\right]$$
$$+\frac{\lambda_i(t)}{\gamma_i b^2(t,y)}h_i'(1-p) \tag{3.16}$$

is a Nash equilibrium with $p\in(0,1)$, *and* $V_i(t,x,y)=x+D_i(t,y)$ *is the equilibrium value function of Agent* i.

From (3.16), it is evident that the equilibrium distribution of Agent i is uniquely determined by his own Choquet regularizer, h_i', and remains independent of his opponent's regularizer, h_j. Furthermore, (3.11) and (3.13) show that while the mean of Agent i's distribution depends on both his own parameters and those of his opponent, the variance is solely determined by his own parameters, specifically λ_i, h_i, and γ_i. These insights align with intuitive expectations in the context of RL. Although an opponent's risk tolerance, sensitivity, or strategic decisions can influence the expected outcomes of the decision-making process, the degree of exploration, as reflected by variance, is solely a function of the agent's intrinsic characteristics. Additionally, (3.11) shows that larger value of λ_i indicates a stronger emphasis on exploration, leading to more dispersed exploration around the current position of Agent i. In contrast, an increase in the risk aversion parameter γ_i reflects a more cautious approach, leading to reduced variance in the exploratory strategy.

3.3 Solution to Gauss mean return model

In this subsection, we examine the Gaussian mean return model as a special case of the state process $Y(t)$ shown in Example 1, that is,

$$a(t,y)=r+\sigma y,\ \ b(t,y)=\sigma,\ \ m(t,y)=\iota(Y-y),\ \ v(t,y)=v, \tag{3.17}$$

where r, σ, ι, v, and Y are positive constants. Thus, by (2.2), we have $\theta(t,y)=y$. We formulate the following proposition as a direct consequence of Theorem 2.

Proposition 2. *For the Gauss mean return model,* $p\in(0,1)$, $i,j\in\{1,2\}$ *and* $i\neq j$, *profile* $(\boldsymbol{\Pi}_1^*,\boldsymbol{\Pi}_2^*)$ *with quantile functions*

$$Q_{\Pi_i^*(t)}(p)=\frac{1}{1-k_1k_2}\left[\frac{y}{\sigma}\left(\frac{1}{\gamma_i}+\frac{k_i}{\gamma_j}\right)-\frac{\rho v}{\sigma}\left((a_2^i(t)+k_ia_2^j(t))y+(a_1^i(t)+k_ia_1^j(t))\right)\right]$$
$$+\frac{\lambda_i(t)}{\gamma_i\sigma^2}h_i'(1-p), \tag{3.18}$$

is a Nash equilibrium. Moreover, the corresponding equilibrium value function V_i *and* g_i *have the following form*

$$V_i(t,\hat{x}_i,y)=\hat{x}_i+\frac{1}{2}b_2^i(t)y^2+b_1^i(t)y+b_0^i(t), \tag{3.19}$$

and

$$g_i(t,\hat{x}_i,y)=\hat{x}_i+\frac{1}{2}a_2^i(t)y^2+a_1^i(t)y+a_0^i(t), \tag{3.20}$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

where $a_n^i(t)$, $b_n^i(t)$, $n = 0, 1, 2$, are continuously differentiable functions defined as

$$\begin{cases} a_0^i(t) = \dfrac{\iota^2 Y^2}{\gamma_i(\iota+\rho v)^2}\left(T - t + \dfrac{1-e^{-2(\iota+\rho v)(T-t)}}{2(\iota+\rho v)} - \dfrac{2(1-e^{-(\iota+\rho v)(T-t)})}{(\iota+\rho v)}\right) \\ \qquad + \dfrac{v^2}{2\gamma_i(\iota+\rho v)}\left(T - t - \dfrac{1-e^{-2(\iota+\rho v)(T-t)}}{2(\iota+\rho v)}\right), \\ a_1^i(t) = \dfrac{\iota Y}{\gamma_i(\iota+\rho v)^2}[1-e^{-(\iota+\rho v)(T-t)}]^2, \\ a_2^i(t) = \dfrac{1}{\gamma_i(\iota+\rho v)}[1-e^{-2(\iota+\rho v)(T-t)}], \end{cases} \tag{3.21}$$

and

$$\begin{cases} {b_2^i}'(t) = 2\iota b_2^i(t) + \gamma_i(1-\rho^2)v^2 a_2^i(t)^2 + 2\rho v a_2^i(t) - \dfrac{1}{\gamma_i}, \\ {b_1^i}'(t) = \iota b_1^i(t) - \iota Y b_2^i(t) + \gamma_i(1-\rho^2)v^2 a_2^i(t) a_1^i(t) + \rho v a_1^i(t), \\ {b_0^i}'(t) = -\iota Y b_1^i(t) - \dfrac{v^2}{2} b_2^i(t) + \dfrac{\gamma_i(1-\rho^2)v^2}{2} a_1^i(t)^2 \\ \qquad + \dfrac{\gamma_i k_i^2 \sigma^2}{2}\sigma_j(t)^2 - \dfrac{\lambda_i^2(t)\|h_i'\|_2^2}{2\gamma_i\sigma^2}, \end{cases} \tag{3.22}$$

with $b_0^i(T) = b_1^i(T) = b_2^i(T) = 0$.

Proof. For the Gauss mean return model, (3.14) can be simplified as

$$\frac{\partial d_i(t,y)}{\partial t} + [\iota(Y-y) - \rho v y]\frac{\partial d_i(t,y)}{\partial y} + \frac{v^2}{2}\frac{\partial^2 d_i(t,y)}{\partial y^2} + \frac{y^2}{\gamma_i} = 0. \tag{3.23}$$

By letting $d_i(t,y) = \frac{1}{2}a_2^i(t)y^2 + a_1^i(t)y + a_0^i(t)$ and substituting it into (3.23), we obtain

$$\begin{cases} {a_2^i}'(t) = 2a_2^i(t)(\iota+\rho v) - \dfrac{2}{\gamma_i}, & a_2^i(T) = 0, \\ {a_1^i}'(t) = a_1^i(t)(\iota+\rho v) - a_2^i(t)\iota Y, & a_1^i(T) = 0, \\ {a_0^i}'(t) = -a_1^i(t)\iota Y - \dfrac{v^2}{2}a_2^i(t), & a_0^i(T) = 0. \end{cases} \tag{3.24}$$

It can be shown that (3.21) is the solution to (3.24). By substituting d_i into (3.16), we derive (3.18). Consequently, $(\mathbf{\Pi}_1^*, \mathbf{\Pi}_2^*)$ is indeed a Nash equilibrium. Similarly, by simplifying (3.15), we can get $D_i(t,y) = \frac{1}{2}b_2^i(t)y^2 + b_1^i(t)y + b_0^i(t)$ with $b_n^i(t)$, $n = 0, 1, 2$, given by (3.22). □

By setting $y = (a-r)/\sigma$, $\iota = 0$ and $v = 0$ in (3.17), the price dynamics (2.1) reduces to the classical Black-Scholes model. In this case, the market becomes complete and the corresponding results are straightforward, as stated below.

Corollary 1. *In the Black-Scholes model, for $p \in (0, 1)$, $i, j \in \{1, 2\}$ and $i \neq j$, profile $(\mathbf{\Pi}_1^*, \mathbf{\Pi}_2^*)$ with quantile functions*

$$Q_{\Pi_i^*(t)}(p) = \frac{1}{1-k_1k_2}\left[\frac{a-r}{\sigma^2}\left(\frac{1}{\gamma_i} + \frac{k_i}{\gamma_j}\right)\right] + \frac{\lambda_i(t)}{\gamma_i\sigma^2}h_i'(1-p) \tag{3.25}$$

is a Nash equilibrium.

In the Black-Scholes model, the influence of k_1, k_2, γ_1, and γ_2 on the equilibrium strategies is clearly reflected in (3.25), aligning with the properties of Gauss mean return model discussed in Section 6.

4. Policy iteration

In this section, we employ the policy iteration method to find equilibrium strategies in two steps. For $i,j \in \{1,2\}$ and $i \neq j$, we first fix $\mathbf{\Pi}_j$ and estimate the associated value function $V_i^{\mathbf{\Pi}_i}$ given a policy $\mathbf{\Pi}_i$. Then we update the previous policy $\mathbf{\Pi}_i$ to a new one $\tilde{\mathbf{\Pi}}_i$ based on the obtained value function $V_i^{\mathbf{\Pi}_i}$. Despite the learning process not leading to a monotone iteration algorithm due to the "optimality" is in the sense of equilibrium, we demonstrate that the iterative process converges uniformly to the desired equilibrium policy.

Assuming $\mathbf{\Pi}_j$ is fixed, and letting $\mathbf{\Pi}_i$ be an admissible strategy for $i,j \in \{1,2\}$ with $i \neq j$, we denote the value function under $\mathbf{\Pi}_i$ as $V_i^{\mathbf{\Pi}_i}(t,\hat{x}_i,y) = J_i(t,\hat{x}_i,y;\mathbf{\Pi}_i,\mathbf{\Pi}_j)$. Similarly, we define $g_i^{\mathbf{\Pi}_i}(t,\hat{x}_i,y) = \mathbb{E}_t[\hat{X}_i^{\Pi_i,\Pi_j}(T)]$. According to Björk *et al*. (2017), the functions $V_i^{\mathbf{\Pi}_i}$ and $g_i^{\mathbf{\Pi}_i}$ satisfy the following equations

$$\mathscr{L}^{\Pi_i,\Pi_j} V_i^{\mathbf{\Pi}_i} - \frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i 2} + \gamma_i g_i^{\mathbf{\Pi}_i} \mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i} + \lambda_i(t)\Phi_{h_i}(\Pi_i) = 0, \tag{4.1}$$

and

$$\mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i}(t,\hat{x}_i,y) = 0, \tag{4.2}$$

with

$$V_i^{\mathbf{\Pi}_i}(T,\hat{x}_i,y) = \hat{x}_i, \quad g_i^{\mathbf{\Pi}_i}(T,\hat{x}_i,y) = \hat{x}_i. \tag{4.3}$$

Theorem 3. *For $p \in (0,1)$, $i,j \in \{1,2\}$, and $i \neq j$, with $\mathbf{\Pi}_j$ fixed, let $\mathbf{\Pi}_i^0$ be the initial policy of Agent i with quantile function given by*

$$Q_{\Pi_i^0(t)}(p) = \frac{y}{\gamma_i \sigma} + k_i \mu_j(t) - \frac{\rho v}{\sigma}(a_2^{i0}(t) y + a_1^{i0}(t)) + \psi^0 h_i'(1-p). \tag{4.4}$$

Choose one policy

$$\Pi_i(t) \in arg\,max_{\Pi_i \in \mathscr{M}^2}\{\mathscr{L}^{\Pi_i,\Pi_j} V_i^{\mathbf{\Pi}_i^n} - \frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i^n 2} + \gamma_i g_i^{\mathbf{\Pi}_i^n} \mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i^n} + \lambda_i(t)\Phi_{h_i}(\Pi_i)\},$$

and denote this policy as $\mathbf{\Pi}_i^{n+1}$, $n = 0,1,2,\ldots$ Then the following statements holds.

(i) The sequence of updated policies $\mathbf{\Pi}_i^n$ for $n \geqslant 1$ has the quantile function

$$Q_{\Pi_i^n(t)}(p) = \frac{y}{\gamma_i \sigma} + k_i \mu_j(t) - \frac{\rho v}{\sigma}(a_2^{in}(t) y + a_1^{in}(t)) + \frac{\lambda_i(t)}{\gamma_i \sigma^2} h_i'(1-p), \tag{4.5}$$

where a_1^{in} and a_2^{in} satisfy

$$\begin{cases} a_2^{in\prime}(t) = 2\iota a_2^{in}(t) + 2\rho v a_2^{in-1}(t) - \dfrac{2}{\gamma_i}, & a_2^{in}(T) = 0, \\ a_1^{in\prime}(t) = \iota a_1^{in}(t) + \rho v a_1^{in-1}(t) - a_2^{in}(t)\iota Y, & a_1^{in}(T) = 0. \end{cases} \tag{4.6}$$

(ii) As $n \to \infty$, $a_1^{in}(t)$ and a_2^{in} uniformly converge to a_1^i and a_2^i in (3.21), respectively.

Proof.

(i) Note that $\mathbf{\Pi}_i^0$ satisfies

$$\mathscr{L}^{\Pi_i,\Pi_j} V_i^{\mathbf{\Pi}_i^0} - \frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j}(g_i^{\mathbf{\Pi}_i^0})^2 + \gamma_i g_i^{\mathbf{\Pi}_i^0} \mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i^0} + \lambda_i(t)\Phi_{h_i}(\Pi_i) = 0,$$

and

$$\mathscr{L}^{\Pi_i,\Pi_j} g_i^{\mathbf{\Pi}_i^0}(t,\hat{x}_i,y) = 0. \tag{4.7}$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

Consider $V_i^{\boldsymbol{\Pi}_i^0}(t,\hat{x}_i,y)=\hat{x}_i+D_i^{\boldsymbol{\Pi}_i^0}(t,y)$ and $g_i^{\boldsymbol{\Pi}_i^0}(t,\hat{x}_i,y)=\hat{x}_i+d_i^{\boldsymbol{\Pi}_i^0}(t,y)$. Substituting $g_i^{\boldsymbol{\Pi}_i^0}$ into (4.7), we get

$$\frac{\partial d_i^{\boldsymbol{\Pi}_i^0}(t,y)}{\partial t}+\iota(Y-y)\frac{\partial d_i^{\boldsymbol{\Pi}_i^0}(t,y)}{\partial y}+\frac{v^2}{2}\frac{\partial^2 d_i^{\boldsymbol{\Pi}_i^0}(t,y)}{\partial y^2}+\frac{y^2}{\gamma_i}-\rho v(a_2^{i0}(t)y^2+a_1^{i0}(t)y)=0. \tag{4.8}$$

Assuming $d_i^{\boldsymbol{\Pi}_i^0}(t,y)=\frac{1}{2}a_2^{i1}(t)y^2+a_1^{i1}(t)y+a_0^{i1}(t)$ and substituting it into (4.8), we get

$$\begin{cases} {a_2^{i1}}'(t)=2\iota a_2^{i1}(t)+2\rho v a_2^{i0}(t)-\dfrac{2}{\gamma_i}, & a_2^{i1}(T)=0,\\ {a_1^{i1}}'(t)=\iota a_1^{i1}(t)+\rho v a_1^{i0}(t)-a_2^{i1}(t)\iota Y, & a_1^{i1}(T)=0,\\ {a_0^{i1}}'(t)=-a_1^{i1}(t)\iota Y-\dfrac{v^2}{2}a_2^{i1}(t), & a_0^{i1}(T)=0. \end{cases} \tag{4.9}$$

By policy iteration, we know that

$$\Pi_i^1(t)\in\arg\max_{\Pi_i\in\mathscr{M}^2}\left\{\mathscr{L}^{\Pi_i,\Pi_j}V_i^{\boldsymbol{\Pi}_i^0}-\frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j}{g_i^{\boldsymbol{\Pi}_i^0}}^2+\gamma_i g_i^{\boldsymbol{\Pi}_i^0}\mathscr{L}^{\Pi_i,\Pi_j}g_i^{\boldsymbol{\Pi}_i^0}+\lambda_i(t)\Phi_{h_i}(\Pi_i)\right\}.$$

By the first-order conditions, we have

$$\mu_i^1(t)=\frac{y}{\gamma_i\sigma}+k_i\mu_j(t)-\frac{\rho v}{\sigma}(a_2^{i1}(t)y+a_1^{i1}(t)),\ \text{and}\ \sigma_i^1(t)=\frac{\lambda_i(t)\|h_i'\|_2}{\gamma_i\sigma^2}. \tag{4.10}$$

Repeating the above procedure, we then get (4.5) and (4.6).

(ii) Denote $M=\sup_{t\in[0,T]}|a_2^i(t)-a_2^{i0}(t)|$, $m=\sup_{t\in[0,T]}|a_1^i(t)-a_1^{i0}(t)|$, $\Delta_{k+1}(t)=a_2^i(t)-a_2^{i(k+1)}(t)$ and $\delta_{k+1}(t)=a_1^i(t)-a_1^{i(k+1)}(t)$. We claim that

$$|\Delta_n(t)|\leqslant\frac{[2\rho v(T-t)]^n}{n!}M. \tag{4.11}$$

The case for $n=0$ is trivial. By induction, we assume that the inequality holds for $n=K$. Then it follows from (3.24) and (4.6) that $\Delta_{k+1}(t)$ satisfies

$$\Delta_{k+1}'(t)=2\iota\Delta_{k+1}(t)+2\rho v\Delta_k(t),\qquad \Delta_{k+1}(T)=0.$$

Solving this differential equation, we obtain $\Delta_{k+1}(t)=-\int_t^T 2\rho v e^{2\iota(t-s)}\Delta_k(s)\mathrm{d}s$. Consequently,

$$|\Delta_{k+1}(t)|\leqslant\int_t^T 2\rho v|\Delta_k(s)|\mathrm{d}s\leqslant\int_t^T 2\rho v\frac{[2\rho v(T-s)]^k}{k!}M\mathrm{d}s=\frac{[2\rho v(T-t)]^{k+1}}{(k+1)!}M.$$

Thus, (4.8) holds. Similarly, we can prove by induction that

$$|\delta_n(t)|\leqslant\frac{[\rho v(T-t)]^n}{n!}m+\frac{\iota Y}{\rho v}\frac{[2\rho v(T-t)]^{n+1}}{(n+1)!}M.$$

Thus, $a_1^{in}(t)$ and a_2^{in} uniformly converge to a_1^i and a_2^i as $n\to\infty$, respectively. □

Theorem 3 shows that the iteration does not change the form of the policy (see (4.5)), and thus, it suffices to parameterize the iterative policy through two deterministic functions $(a_1^{(in)}(t),a_2^{(in)}(t))$. In particular, when the initial policy chosen as the form of the equilibrium policy in Proposition 2, our algorithm is guaranteed to uniformly converge to the equilibrium policy.

The next result guarantees the convergence of policies as the two agents iterate simultaneously.

Theorem 4. *For an initial profile* $(\boldsymbol{\Pi}_1^0,\boldsymbol{\Pi}_2^0)$*, assume that two agents iterate simultaneously by* (4.5) *and the updated sequence is defined by* $(\boldsymbol{\Pi}_1^n,\boldsymbol{\Pi}_2^n)$*,* $n=0,1,2,\ldots$ *Then for* $p\in(0,1)$*,* $i,j\in\{1,2\}$ *and* $i\neq j$*,* $Q_{\Pi_i^n(t)}(p)$ *converges uniformly to* $Q_{\Pi_i^*(t)}(p)$ *of* (3.18) *as* $n\to\infty$*.*

Proof. We just need to prove that the mean and variance of Π_i converge to the mean and variance of Π_i^*. The convergence of variance is obvious according to the proof of Theorem 3(i) and (4.10). Let $\mu_i^n(t)$ be the mean of $\Pi_i^n(t)$. Based on (3.10), we have

$$\begin{bmatrix}\mu_1^{n+1}(t)\\ \mu_2^{n+1}(t)\end{bmatrix}=\begin{bmatrix}0 & k_1\\ k_2 & 0\end{bmatrix}\begin{bmatrix}\mu_1^{n}(t)\\ \mu_2^{n}(t)\end{bmatrix}+\begin{bmatrix}\dfrac{y}{\gamma_1\sigma}-\dfrac{\rho v}{\sigma}(a_2^1(t)y+a_1^1(t))\\ \dfrac{y}{\gamma_2\sigma}-\dfrac{\rho v}{\sigma}(a_2^2(t)y+a_1^2(t))\end{bmatrix}.$$

Consider the normed space $\mathbb{R}^2$ with $\|\cdot\|$ defined as $\|\vec{x}\|=\max\{x_1,x_2\}$ for $\vec{x}=[x_1,x_2]'\in\mathbb{R}^2$. It is well known that $(\mathbb{R}^2,\|\cdot\|)$ is a Banach space. Define

$$f(\vec{x})=\begin{bmatrix}0 & k_1\\ k_2 & 0\end{bmatrix}\vec{x}+\begin{bmatrix}\dfrac{y}{\gamma_1\sigma}-\dfrac{\rho v}{\sigma}(a_2^1(t)y+a_1^1(t))\\ \dfrac{y}{\gamma_2\sigma}-\dfrac{\rho v}{\sigma}(a_2^2(t)y+a_1^2(t))\end{bmatrix}.$$

We have $\|f(\vec{x})-f(\vec{y})\|\leqslant\max\{k_1,k_2\}\|\vec{x}-\vec{y}\|$, establishing f as a contraction mapping with a unique fixed point. By (3.12), the fixed point is precisely $[\mu_1^*(t),\mu_2^*(t)]'$. Thus, the mean of $\Pi_i^n(t)$ also converges. Moreover, let M' be an uniformly upper bound of $|\mu_i^0(t)-\mu_i^*(t)|$. Then, we have

$$\begin{aligned}\left\|\begin{bmatrix}\mu_1^n(t)\\ \mu_2^n(t)\end{bmatrix}-\begin{bmatrix}\mu_1^*(t)\\ \mu_2^*(t)\end{bmatrix}\right\|&=\left\|\begin{bmatrix}0 & k_1\\ k_2 & 0\end{bmatrix}\begin{bmatrix}\mu_1^{n-1}(t)-\mu_1^*(t)\\ \mu_2^{n-1}(t)-\mu_2^*(t)\end{bmatrix}\right\|\\ &=\left\|\begin{bmatrix}0 & k_1\\ k_2 & 0\end{bmatrix}^n\begin{bmatrix}\mu_1^{0}(t)-\mu_1^*(t)\\ \mu_2^{0}(t)-\mu_2^*(t)\end{bmatrix}\right\|\\ &\leqslant M'(\max\{k_1,k_2\})^n.\end{aligned}$$

Thus, for $p\in(0,1)$, $i,j\in\{1,2\}$ and $i\neq j$, $Q_{\Pi_i^n(t)}(p)$ converges uniformly to $Q_{\Pi_i^*(t)}(p)$ of (3.18) as $n\to\infty$. □

5. RL algorithm design

In this section, we devise an algorithm to learn the Nash equilibrium. As mentioned in the introduction, game scenarios involve multiple agents in the environment, requiring the utilization of multi-agent reinforcement learning algorithms, which inherently introduce greater complexity. Specifically, in single-agent reinforcement learning, a basic assumption is the stability of the environment, wherein the transition probability and reward function remain constant. However, when other intelligent agents are introduced into the environment, this assumption no longer holds true. In a multi-agent context, any change in one agent's strategy can significantly impact other agents, leading to dynamic evolution of the environment with their strategies. Moreover, as the number of agents increases, the complexity of training also escalates. Fortunately, in our model, each agent will only affect the mean of the other agent's strategy. According to (3.12), the difference $\mu_i(t)-k_i\mu_j(t)$ for $i,j\in\{1,2\}$ and $i\neq j$ remains fixed, regardless of how the agents adjust their strategies. Thus, once the difference is known, we can directly apply the learning procedure outlined in Theorem 4. This allows us to reduce the game problem into two independent optimization problems. We then apply the method proposed in Section 4 to learn the difference. Below, we briefly introduce this method.

Assume that the risk-free interest rate r, the risk-aversion coefficient γ_1,γ_2, the sensitivity coefficient k_1,k_2, and the exploration weight λ_1,λ_2 are known. The agents have no prior knowledge of $S(t)$ and $Y(t)$, but can observe the pair $(S(t),Y(t))$ at each time t. In the continuous time setting, we discretize the interval $[0,T]$ into N subintervals with equal length $\Delta t=t_{k+1}-t_k, k=0,1,\ldots,N-1$. Based on (2.1) and (2.3), we get

$$\mathrm{d}X_i^{u_i}(t)=u_i(t)\frac{\mathrm{d}e^{-rt}S(t)}{e^{-rt}S(t)}.$$

Therefore, when the Agent i follows strategy $\Pi_i(t_k)$ at time t_k and samples action $u_i(t_k)$ from $\Pi_i(t_k)$, the discounted wealth at t_{k+1} is

$$X_i^{\Pi_i}(t_{k+1}) \approx X_i^{\Pi_i}(t_k) + u_i(t_k)\frac{e^{-rt_{k+1}}S(t_{k+1}) - e^{-rt_k}S(t_k)}{e^{-rt_k}S(t_k)}.$$

Assume that the policy $\mathbf{\Pi}_j$ of Agent j is fixed, and that both $\Pi_i(t)$ and $\Pi_j(t)$ have density $\pi_i(t)$ and $\pi_j(t)$. The algorithm for learning the difference $\mu_i(t) - k_i\mu_j(t)$ is based on the standard idea of policy evaluation followed by policy update. We do not adopt the approach of generalized policy iteration, in which policy evaluation and policy update interact with each other. There are three main reasons for this. First, Theorem 3 guarantees a convergence rate that is sufficiently fast, making the generalized policy iteration unnecessary. Second, unlike classical reinforcement learning, our equilibrium strategy involves two value functions, V_i and g_i, which complicates the policy evaluation step, particularly in the context of generalized policy iteration. Third, as previously discussed, our policy update does not always yield an improved policy, which further limits the applicability of generalized policy iteration.

We represent the policy $\mathbf{\Pi}_i$ of Agent i using its quantile function. By Theorem 3, the quantile function can be defined as

$$Q_{\Pi_i^{\Psi}(t)}(p) = k_i\mu_j(t) + \psi_0\frac{y}{\gamma_i} - \psi_0\psi_1(a_2^{i0}(t)y + a_1^{i0}(t)) + \frac{\lambda_i}{\gamma_i\psi_2^2}h_i'(1-p), \tag{5.1}$$

where $\Psi = (\psi_0, \psi_1, \psi_2) \in \mathbb{R}^3$ denotes the set of parameters to be learned, and a_1^{i0} and a_2^{i0} are functions that will be updated according to Theorem 3(i). The parameters Ψ can be initialized either randomly or using predetermined constants, and a_1^{i0} and a_2^{i0} can be initialized as zero functions.

For policy evaluation procedure, based on Proposition 2, we parameterize V_i and g_i in the form of (3.19) and (3.20), respectively, as follows

$$\begin{aligned} V_i^{\Theta}(t, \hat{x}_i, y) &= \hat{x}_i + \frac{1}{2}p(\theta_i^{V,2}, T-t)y^2 + p(\theta_i^{V,1}, T-t)y + p(\theta_i^{V,0}, T-t), \\ g_i^{\Theta}(t, \hat{x}_i, y) &= \hat{x}_i + \frac{1}{2}p(\theta_i^{g,2}, T-t)y^2 + p(\theta_i^{g,1}, T-t)y + p(\theta_i^{g,0}, T-t), \end{aligned} \tag{5.2}$$

where $p(\theta, t)$ is a suitable parametric function with coefficient vector $\theta \in \mathbb{R}^d$. The full parameter set is denoted by $\Theta = (\theta_i^{V,0}, \theta_i^{V,1}, \theta_i^{V,2}, \theta_i^{g,0}, \theta_i^{g,1}, \theta_i^{g,2}) \in \mathbb{R}^{6d}$. A typical choice for $p(\theta, t)$ is a linear combination of the first d terms of a basis expansion, such as a truncated Taylor or Fourier series, where θ denotes the corresponding coefficients. Since $g^{\mathbf{\Pi}_i}$ satisfies $\mathscr{L}^{\Pi_i,\Pi_j}g_i^{\mathbf{\Pi}_i}(t, \hat{x}_i, y) = 0$ with terminal condition $g_i^{\mathbf{\Pi}_i}(T, \hat{x}_i, y) = \hat{x}_i$ for any $\mathbf{\Pi}_i$, it can be interpreted via the Feynman-Kac formula as the value function of a time-consistent optimal problem. This allows it to be evaluated using continuous-time reinforcement learning. Theorem 3 in Jia and Zhou (2022a) and Theorem 4 in Jia and Zhou (2022b) show that $g^{\mathbf{\Pi}_i}$ can be estimated by minimizing the martingale loss function

$$\mathrm{ML}_g(\Theta) := \frac{1}{2}\mathbb{E}\left[\sum_{k=0}^{N-1}\left(\hat{X}_i(T) - g_i^{\Theta}(t_k, \hat{X}_i(t_k), Y(t_k))\right)^2\Delta t\right],$$

which corresponds to the continuous-time analogue of the *Monte Carlo* policy evaluation with function approximation (e.g., Sutton and Barto, 2018). Then we can use the stochastic gradient descent method to minimize $\mathrm{ML}_g(\Theta)$, which is a standard method in reinforcement learning. The gradient of $\mathrm{ML}_g(\Theta)$ with respect to Θ can be computed as

$$\nabla_g\Theta = -\sum_{k=0}^{N-1}\left(\hat{X}_i(T) - g_i^{\Theta}(t_k, \hat{X}_i(t_k), Y(t_k))\right)\frac{\partial g_i^{\Theta}}{\partial\Theta}\Delta t, \tag{5.3}$$

and the parameters Θ are updated according to

$$\Theta_{n+1} = \Theta_n - \alpha_g\nabla_g\Theta.$$

Once g_i is evaluated, we can proceed to evaluate V_i, which satisfies

$$\mathscr{L}^{\Pi_i,\Pi_j} V_i(t,\hat{x}_i,y) + R(t,\hat{x}_i,y) = 0, \quad V_i(T,\hat{x}_i,y) = \hat{x}_i, \tag{5.4}$$

where the residual term is given by

$$R(t,\hat{x}_i,y) = -\frac{\gamma_i}{2}\mathscr{L}^{\Pi_i,\Pi_j} g_i^2 + \gamma_i g_i \mathscr{L}^{\Pi_i,\Pi_j} g_i + \lambda_i(t)\Phi_{h_i}(\Pi_i).$$

The structure of (5.4), together with the Feynman-Kac formula, implies that V_i can also be interpreted as the value function of a time-consistent optimal control problem. Consequently, V_i can be evaluated analogously to g_i, by minimizing a martingale loss function

$$\mathrm{ML}_V(\Theta) := \frac{1}{2}\mathbb{E}\left[\sum_{k=0}^{N-1}\left(\hat{X}_i(T) - V_i^{\Theta}(t_k,\hat{X}_i(t_k),Y(t_k)) + \sum_{l=k}^{N-1} R(t_l,\hat{X}_i(t_l),Y(t_l))\Delta t\right)^2 \Delta t\right],$$

and the gradient of $\mathrm{ML}_V(\Theta)$ with respect to Θ can be computed as

$$\nabla_V\Theta = -\sum_{k=0}^{N-1}\left(\hat{X}_i(T) - V_i^{\Theta}(t_k,\hat{X}_i(t_k),Y(t_k)) + \sum_{l=k}^{N-1} R(t_l,\hat{X}_i(t_l),Y(t_l))\Delta t\right)\frac{\partial V_i^{\Theta}}{\partial\Theta}\Delta t. \tag{5.5}$$

Then Θ can be updated as

$$\Theta_{n+1} = \Theta_n - \alpha_V \nabla_V \Theta.$$

By applying Itô's formula, for any $\varphi \in C^{1,2,2}$ and strategy Π_i of Agent i, we have

$$\begin{aligned}&\mathbb{E}_t[\varphi(t+\Delta t,\hat{X}_i^{\Pi_i,\Pi_j}(t+\Delta t),Y(t+\Delta t))] - \varphi(t,\hat{X}_i^{\Pi_i,\Pi_j}(t),Y(t))\\ &= \mathbb{E}_t\int_t^{t+\Delta t}\mathscr{L}^{\Pi_i,\Pi_j}\varphi(s,\hat{X}_i^{\Pi_i,\Pi_j}(s),Y(s))\mathrm{d}s,\end{aligned}$$

and thus

$$\mathscr{L}^{\Pi_i,\Pi_j}\varphi(t,\hat{X}_i^{\Pi_i,\Pi_j}(t),Y(t)) \approx \frac{\varphi(t+\Delta t,\hat{X}_i^{\Pi_i,\Pi_j}(t+\Delta t),Y(t+\Delta t)) - \varphi(t,\hat{X}_i^{\Pi_i,\Pi_j}(t),Y(t))}{\Delta t}. \tag{5.6}$$

We can use (5.6) to approximate $R(t,\hat{x}_i,y)$ by replacing φ with V_i, g_i and g_i^2.

For policy update procedure, we want to maximize the following function

$$L_i(\Psi;t,\hat{x},y) := \mathscr{L}^{\Pi_i^{\Psi}(t),\Pi_j} V_i^{\Theta} - \frac{\gamma_i}{2}\mathscr{L}^{\Pi_i^{\Psi}(t),\Pi_j} g_i^{\Theta 2} + \gamma_i g_i^{\Theta}\mathscr{L}^{\Pi_i^{\Psi}(t),\Pi_j} g_i^{\Theta} + \lambda_i\Phi_{h_i}(\Pi_i^{\Psi}(t)) \tag{5.7}$$

for all possible $t,\hat{x},y$. By the proof of Theorem 3, we only need to maximize $L_i(\Psi;t,\hat{x},y)$ at initial state $(t_0,\hat{x}_0,y_0)$. First, we can update policy (5.1) to

$$Q_{\Pi_i^{\Psi}(t)}(p) = k_i\mu_j(t) + \psi_0\frac{y}{\gamma_i} - \psi_0\psi_1(a_2^{i1}(t)y + a_1^{i1}(t)) + \frac{\lambda_i(t)}{\gamma_i\psi_2^2}h_i'(1-p), \tag{5.8}$$

where $a_2^{i1}(t)y + a_1^{i1}(t)$ comes from the derivative of evaluated g^{Θ} with respect to y. Then we only need to maximize $L_i(\Psi;t_0,\hat{x}_0,y_0)$ with respect to Ψ. The maximization of $L_i(\Psi;t_0,\hat{x}_0,y_0)$ can also be achieved by stochastic gradient ascent method. Since

$$\mathscr{L}^{\Pi_i,\Pi_j}\varphi(t,\hat{x},y) = \int_R\int_R \mathscr{L}^{u_i,u_j}\varphi(t,\hat{x},y)\pi_j(u_j)\pi_i(u_i)\mathrm{d}u_j\mathrm{d}u_i,$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

then we have

$$\begin{aligned}\frac{dL_i}{d\Psi} &= \int_R\int_R\Bigg[\left(\mathscr{L}^{u_i,u_j}V_i^{\Theta} - \frac{\gamma_i}{2}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \gamma_i g_i^{\Theta}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \lambda_i\Phi_{h_i}(\Pi_i^{\Psi})\right)\pi_j(u_j)\frac{\partial\pi_i^{\Psi}(u_i)}{\partial\Psi}\\ &\quad + \lambda_i\frac{\partial\Phi_{h_i}(\Pi_i^{\Psi})}{\partial\Psi}\pi_j(u_j)\pi_i(u_i)\Bigg]\mathrm{d}u_j\mathrm{d}u_i.\\ &= \int_R\int_R\Bigg[\left(\mathscr{L}^{u_i,u_j}V_i^{\Theta} - \frac{\gamma_i}{2}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \gamma_i g_i^{\Theta}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \lambda_i\Phi_{h_i}(\Pi_i^{\Psi})\right)\pi_j(u_j)\pi_i(u_i)\frac{\frac{\partial\pi_i^{\Psi}(u_i)}{\partial\Psi}}{\pi_i(u_i)}\\ &\quad + \lambda_i\frac{\partial\Phi_{h_i}(\Pi_i^{\Psi})}{\partial\Psi}\pi_j(u_j)\pi_i(u_i)\Bigg]\mathrm{d}u_j\mathrm{d}u_i.\\ &= \int_R\int_R\Bigg[\frac{\partial\log\pi_i^{\Psi}(u_i)}{\partial\Psi}\left(\mathscr{L}^{u_i,u_j}V_i^{\Theta} - \frac{\gamma_i}{2}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \gamma_i g_i^{\Theta}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \lambda_i\Phi_{h_i}(\Pi_i^{\Psi})\right)\\ &\quad + \lambda_i\frac{\partial\Phi_{h_i}(\Pi_i^{\Psi})}{\partial\Psi}\Bigg]\pi_j(u_j)\pi_i(u_i)\mathrm{d}u_j\mathrm{d}u_i.\end{aligned}$$

Thus, the gradient of $L_i(\Psi)$ with respect to Ψ can be computed as

$$\nabla\Psi = \frac{\partial\log\pi_i^{\Psi}(u_i)}{\partial\Psi}\left(\mathscr{L}^{u_i,u_j}V_i^{\Theta} - \frac{\gamma_i}{2}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \gamma_i g_i^{\Theta}\mathscr{L}^{u_i,u_j}g_i^{\Theta} + \lambda_i\Phi_{h_i}(\Pi_i^{\Psi})\right) + \lambda_i\frac{\partial\Phi_{h_i}(\Pi_i^{\Psi})}{\partial\Psi}, \tag{5.9}$$

and the infinitesimal generator $\mathscr{L}^{u_i,u_j}$ can be approximated by samples. Note that if the distribution Π^{Ψ} is supported on $[S_{\min}^{\Psi}, S_{\max}^{\Psi}]$ that depends on the parameter Ψ, then the gradient computation must also account for the derivatives of the interval endpoints, $S_{\min}^{\Psi}$ and $S_{\max}^{\Psi}$, with respect to Ψ. Then we can update Ψ as

$$\Psi_{n+1} = \Psi_n + \alpha_{\Psi}\nabla\Psi.$$

By repeating policy evaluation and policy update, we can obtain the final policy. The complete procedure is summarized in Algorithm 1.

6. Numerical results

Given the variety of Choquet regularizers available, it is possible to select different regularizers for each agent. In Equation (2.13), h′(x) represents the "probability weight" assigned to x when calculating the (nonlinear) Choquet expectation (see, e.g., Gilboa and Schmeidler, 1989 and Quiggin, 1982). Consequently, the choice of the distortion function h can directly influence the agent's attitude toward risk. As shown by Han *et al*. (2023), Choquet regularizers can generate several widely used exploratory samplers, such as the ε-greedy strategy, exponential, uniform, and Gaussian. Below, we assume that the agents adopt different Choquet regularizers, resulting in their optimal exploration distributions being normal and exponential, respectively.

Assume that Agent 1 applies the the Choquet regularizer

$$\Phi_{h_1}(\Pi_1) = \int_0^1 Q_{\Pi_1}(p)z(p)\mathrm{d}p, \tag{6.1}$$

where z is the quantile function of a standard normal distribution, yielding $h_1(p) = \int_0^p z(1-s)\mathrm{d}s$ with $p \in [0, 1]$. Further, Agent 2 uses the Choquet regularizer

$$\Phi_{h_2}(\Pi_2) = -\int_0^1 Q_{\Pi_2}(1-p)(\log(p)+1)\mathrm{d}p. \tag{6.2}$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

Algorithm 1

Input: initial wealth x_1, x_2, risk-free interest rate r, exploration weight λ_1, λ_2, the parameters $(\sigma, \iota, Y, \nu, \rho)$ of Market, investment horizon T, time step Δt, number of time grids K, learning rates α, number of samples M, number of iterations N, risk-aversion coefficient γ_1, γ_2, sensitivity coefficient k_1, k_2 and a simulator of the market called *Market*.

Learning procedure:

```
Initialize $\Theta$, $\Psi$ and $a_1^{i0}(t)$, $a_2^{i0}(t)$.
Calculate and store wealth trajectories for each sample.
  for $n = 1$ to $N$ do

      Policy evaluation:
      for $m = 1$ to $M$ do
          Take the $m$th sample and its corresponding wealth trajectory.
          Calculate $\nabla_g \Theta$ by (5.3).
          Update $\Theta$ as $\Theta \leftarrow \Theta - \alpha_g \nabla_g \Theta$.
      end for
      for $m = 1$ to $M$ do
          Take the $m$th sample and its corresponding wealth trajectory.
          Calculate $\nabla_V \Theta$ by (5.5).
          Update $\Theta$ as $\Theta \leftarrow \Theta - \alpha_V \nabla_V \Theta$.
      end for

      Policy update:
      Update $a_1^{i0}(t)$ and $a_2^{i0}(t)$ according to (5.8).
      for $m = 1$ to $M$ do
          Sample from (5.8) and calculate the wealth trajectory.
          Calculate $\nabla \Psi$ by (5.9).
          Update $\Psi$ as $\Psi \leftarrow \Psi + \alpha_\Psi \nabla \Psi$.
      end for

  end for
```

It is known as the *cumulative residual entropy* (e.g., Rao *et al*., 2004 and Hu and Chen, 2020) and $h_2(p) = -p \log p$ with $p \in [0, 1]$. For Gauss mean return model, let

$$\mu_i^*(t) = \frac{1}{1 - k_1 k_2} \left[\frac{y}{\sigma} \left(\frac{1}{\gamma_i} + \frac{k_i}{\gamma_j} \right) - \frac{\rho \nu}{\sigma} \left((a_2^i(t) + k_i a_2^j(t)) y + (a_1^i(t) + k_i a_1^j(t)) \right) \right],$$

where $a_n^i(t)$, $n = 1, 2$, are given by (3.21). Based on (3.16), the equilibrium distribution Π_1^* is a normal distribution given as

$$\Pi_1^*(\cdot; t) = \mathcal{N} \left(\mu_1^*(t), \frac{\lambda_1^2(t)}{\gamma_1^2 \sigma^4} \right),$$

and the equilibrium distribution Π_2^* is an exponential distribution given as

$$\Pi_2^*(u; t) = 1 - \exp \left[\frac{\gamma_2 \sigma^2}{\lambda_2(t)} (\mu_2(t) - u) - 1 \right].$$

Table 1. *Parameter values used in the model.*

ρ	r	σ	ι	v	Y	γ_1	γ_2	k_1	k_2	λ_0	T
−0.93	0.017	0.15	0.27	0.065	0.273	1	2	0.05	0.1	0.01	1

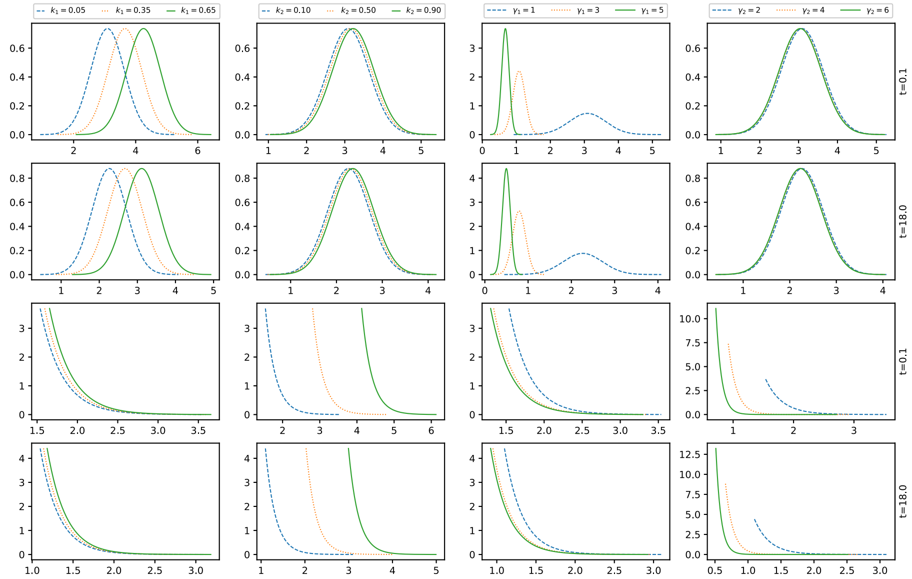

Figure 1. *The effects of t, k_1, k_2, γ_1, and γ_2 on the Nash equilibrium.*

We first investigate the influence of parameters k_i, γ_i and t on the equilibrium strategies of both agents. We assume that $\lambda_1(t) = \lambda_2(t) = \lambda_0 e^{\lambda_0(T-t)}$. Unless otherwise specified, the parameters in (3.16) are set as in Table 1.[1]

In Figure 1, we set $T = 20$ to make the differences between the plotted curves visually distinguishable. The first and second rows display the density functions for Agent 1 at $t = 0.1$ and $t = 18$, respectively, while the third and fourth rows correspond to Agent 2's density functions. For clarity, we focus on the characteristics of the parameters for Agent 1, as Agent 2's performance with respect to these parameters is similar. The key observations are as follows.

(i) As k_1 increases, Agent 1 tends to adopt riskier strategies, leading to a higher mean investment in risky assets. This suggests that greater sensitivity to the opponent's performance enhances Agent 1's motivation to outperform.

(ii) As Agent 1's risk aversion parameter γ_1 rises, the mean of the equilibrium distribution decreases. This indicates that higher levels of risk aversion prompt Agent 1 to adopt more cautious strategies, reducing their expected investment in risky assets.

(iii) As Agent 2 becomes more sensitive to Agent 1's behavior (i.e., as k_2 increases), Agent 1 tends to adopt riskier strategies to increase the likelihood of achieving higher returns. This suggests that Agent 2's increased sensitivity further motivates Agent 1 to excel. Economically, when relative wealth is taken into account, a larger k_2 leads Agent 2 to hold more risky assets, which in turn compels Agent 1 to increase its holdings of risky assets as well in order to maintain the current position and avoid falling behind.

[1] For the Gauss mean return model, Wachter (2002) estimates the market parameters, and Dai *et al*. (2023) uses these parameters for their algorithm. In this work, we also adopt these parameters for our analysis.

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

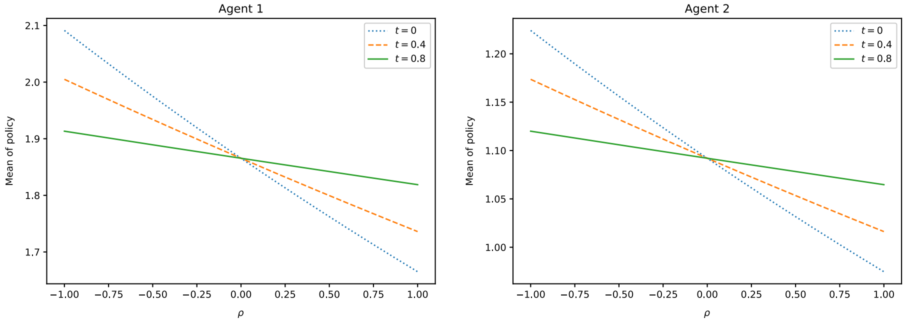

Figure 2. *The effects of t and ρ on the mean of the Nash equilibrium.*

(iv) As Agent 2 becomes more risk-seeking (i.e., as γ_2 decreases), Agent 1 adopts more aggressive strategies, investing a higher mean in risky assets. This adjustment is necessary to prevent Agent 1 from losing market share or competitive advantages if they fail to align their strategy with the increased risks.

(v) Comparing the equilibrium strategies between the first and second rows, the main difference lies in time t. Since a time-decaying temperature parameter $\lambda_1(t)$ is employed, it means that as time progresses, the weight assigned to exploration decreases as time progresses. Consequently, the variance of the equilibrium strategies also decreases, as reflected in (3.11).

It is noteworthy that an agent's own parameters consistently have a greater impact on their strategy than those of their opponent. This outcome is expected. For instance, when the opponent places greater emphasis on the wealth gap, it is the opponent who adjusts their behavior by increasing their own investment. The optimal response of the original decision-maker remains relatively stable, as the opponent's concern for relative performance predominantly affects their own risk-taking behavior, rather than prompting significant changes in the other agent's strategy.

Furthermore, Figure 1 shows that the mean of the equilibrium strategies decreases over time t. This trend can be attributed to the negative value chosen for the parameter ρ. In the following, we investigate how the mean strategy evolves with respect to both ρ and t.

Figure 2 shows that as the correlation ρ increases, the mean of the strategies for both Agent 1 and Agent 2 decrease. This is because positive correlation between the Brownian shocks driving the asset price and the factor process $Y(t)$ increases uncertainty in the asset dynamics, increasing effective risk exposure and making the investor more cautious. However, when risks are negatively correlated ($\rho \in [-1, 0)$), adverse movements in $Y(t)$ tend to offset those in the asset price through the diffusion term, effectively providing a natural hedging effect that reduces risk. As a result, the investor is willing to allocate more on average under negative correlation.

Additionally, for negatively correlated risks, we find that the mean investment decreases as time t increases. This is because as time advances toward the terminal time T, the opportunity to exploit this hedging effect through dynamic rebalancing diminishes. The shorter the remaining horizon, the less effective the negative correlation becomes at mitigating risk over time, leading the investor to reduce the position. Conversely, when risks are positively correlated, the mean investment increases with time t. At early stages, the cumulative risk from positively correlated shocks is higher, prompting caution. However, as t approaches T, the remaining exposure horizon shortens, reducing the impact of correlated shocks on total risk and allowing the investor to increase the allocation. Specifically, when $\rho = 0$, the shocks driving the factor process $Y(t)$ and the asset price are independent. In this case, the mean investment is time-invariant, resulting in identical values at different t, as can be seen directly from (3.18).

Next, we conduct numerical experiments with simulated data to demonstrate our Algorithm 1 for Gauss mean return model and Black-Scholes model. We first emphasize that there are some factors

Table 2. *Parameter settings for the algorithm.*

T	N	γ_1	γ_2	k_1	k_2	λ_1	λ_2
1	250	1	2	0.05	0.1	0.02	0.015

that affect the accuracy of the algorithm. First of all, Theorems 3 and 4 give theoretical convergence, which provides support for algorithm design, but the theoretical results depend on the parameters of the market model and these parameters are actually unknown. Secondly, as mentioned earlier, a very important property in classical reinforcement learning is that the strategy obtained after each update is better than before, but this property no longer holds in time-inconsistent problem. These increases the impact of errors on the convergence in each algorithm iteration, which may affect the convergence to the true equilibrium strategy. Finally, due to the use of maximizing $L_i(\Psi; t_0, \hat{X}_i(t_0), Y(t_0))$ instead of maximizing $L_i(\Psi; t, \hat{x}, y)$ for all possible $t, \hat{x}, y$, each iteration highly relies on the current sample, and different samples may cause the algorithm to converge to different strategies. Meanwhile, similar to classical reinforcement learning, due to the time horizon is finite, each sample only has $N+1$ time points, which is far less than the number required to make the strategy maximizing $L_i(\Psi; t_0, \hat{X}_i(t_0), Y(t_0))$ and maximizing $L_i(\Psi; t, \hat{x}, y)$ for all possible $t, \hat{x}, y$ close enough. In summary, due to various reasons, algorithms for time-inconsistency problems rely more on model settings, especially parameter selection, than the classical reinforcement learning.

We use the stock process parameters detailed in Table 1. Additionally, the other parameter settings for the algorithm are presented in Table 2. For Gauss mean return model, the value function is parameterized as in (5.2) with $p(\theta, t)$ chosen as

$$p(\theta, t) = \theta_1 t^2 + \theta_0 t.$$

The policy is parameterized as (5.1) with suitable initial values selected based on the problem context. From Theorem 3, we know that the optimal values of ψ_0 and ψ_2 are reciprocals of each other, and only once policy update is needed to get the optimal value of ψ_2. Therefore, before starting training, we can perform a pretraining to get optimal ψ_2 to reduce the amount of subsequent training. We examine the behavior of the mean value under Nash equilibrium when Agent 1 adopts a normal distribution and Agent 2 follows an exponential distribution. The corresponding Choquet regularizers are specified in Equations (6.1) and (6.2). Figure 3 presents the mean of the discounted value invested in the risky asset under Nash equilibrium, evaluated at the market state $Y(t) = 0.273$, while Figure 4 illustrates the learning trajectory of ψ_2, which captures the evolution of the variance. By comparing the mean and variance of learned policy with the true policy in Figures 3 and 4, we observed that our experimental results closely approximate the theoretical values. This underscores the effectiveness of our approach in approximating Nash equilibrium solutions. Owing to the asymmetry and heavier tail of the exponential distribution compared to the normal distribution, Figure 4 shows that the trajectory under exponential-distribution-based exploration exhibits greater volatility, though it still converges to the true value. From this standpoint, the normal distribution may appear to yield more stable learning dynamics. However, this interpretation should be viewed with caution, as the regularizers used here capture agents' subjective preferences rather than purely statistical features of the distributions.

For Black-Scholes model, policy evaluation and policy updates can still be implemented using the methods mentioned above. The difference is that the quantile function of the policy $\mathbf{\Pi}_i$ is given by

$$Q_{\Pi_i^{\Psi}(t)}(p) = k_i \mu_j(t) + \psi_0 \frac{y}{\gamma_i} + \frac{\lambda_i}{\gamma_i \psi_1^2} h_i'(1-p),$$

and the value function is parameterized as

$$V_i^{\Theta}(t, \hat{x}_i) = \hat{x}_i + \theta_i^V (T-t) \quad \text{and} \quad g_i^{\Theta}(t, \hat{x}_i) = \hat{x}_i + \theta_i^g (T-t).$$

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

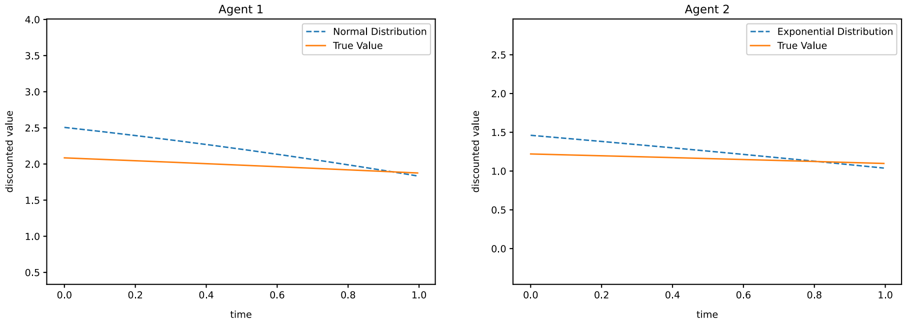

Figure 3. *The mean value of Nash equilibrium.*

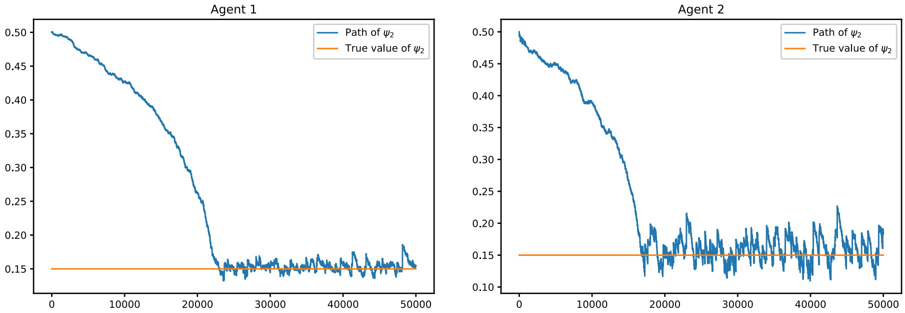

Figure 4. *Convergence of the learned* ψ_2*.*

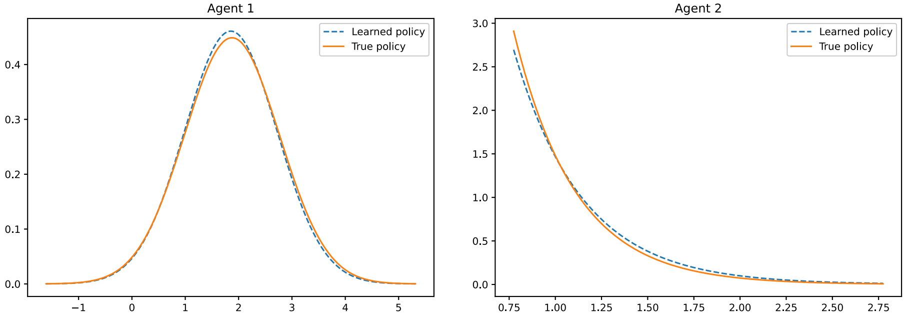

Figure 5. *Learned policy for Black-Scholes model.*

From Corollary 1, we know that the optimal strategy does not change over time and the optimal value of $1/\psi_0$ and ψ_1 is σ. Thus we are actually learning the optimal σ. We plot the density function of the equilibrium policy in Figure 5 when $Y(t) = 0.273$ and $\sigma = 0.15$. We can see that the output of our algorithm closely approximates the theoretical value. Note that the “true” value or policy in Figures 3–5 is computed under the assumption of full knowledge of all model parameters shown in Tables 1 and 2.

7. Conclusion

This paper introduces a framework for continuous-time RL in a competitive market, where two agents consider both their own wealth and their opponent’s wealth under the mean-variance criterion. The Nash equilibrium distributions are derived through dynamic programming, revealing that an agent’s mean of

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

equilibrium distribution for exploration is influenced not only by his own parameters but also by those of his opponent, while the variance of the distribution is solely determined by the agent's own model parameters.

In addition, we investigate the impact of model parameters on the equilibrium strategies, providing valuable insights into decision-making dynamics in competitive financial markets. Furthermore, we design an algorithm to study Nash equilibrium policies, and our experimental results indicate that the output of our algorithm closely approximates the theoretical value.

Notation

$\mathscr{M}$	Set of Borel probability measures or distribution functions on $\mathbb{R}$.
$\mathscr{M}^p$	Subset of $\mathscr{M}$ with finite p-th order moment, $p \in [1, \infty)$.
$S_0(t)$	Price process of the risk-free asset.
$S(t)$	Price process of the risky asset.
$Y(t)$	Process of the state variable representing macroeconomic or systemic risk factors.
$W, \widetilde{W}, \overline{W}_i$	Brownian motions.
ρ	Correlation coefficient between $Y(t)$ and $W(t)$, $\rho \in [-1, 1]$.
γ_i	Risk aversion coefficient of Agent i.
k_i	Sensitivity coefficient of Agent i to the wealth gap with Agent j.
λ_i	Exploration weight of Agent i.
u_i	Control variable of Agent i at some time t.
$\boldsymbol{u}_i$	Control process of Agent i, $\boldsymbol{u}_i = \{u_i(t), 0 \leqslant t \leqslant T\}$.
$X_i^{u_i}(t)$	Discounted wealth process of Agent i under control u_i.
$\hat{X}_i^{u_i,u_j}(t)$	$X_i^{u_i}(t) - k_i X_j^{u_j}$.
Π_i	A distribution randomized from control u_i.
π_i	Density function of Π_i.
$\boldsymbol{\Pi}_i$	Randomized control process of Agent i, $\boldsymbol{\Pi}_i = \{\Pi_i(t), 0 \leqslant t \leqslant T\}$.
$X_i^{\Pi_i}(t)$	Exploratory discounted wealth process of Agent i under randomized control Π_i.
$\hat{X}_i^{\Pi_i,\Pi_j}(t)$	$X_i^{\Pi_i}(t) - k_i X_j^{\Pi_j}(t)$.
$\mathscr{L}^{u_i,u_j}$	Infinitesimal generator of $(\hat{X}_i^{u_i,u_j}(t), Y(t))$.
$\mathscr{L}^{\Pi_i,\Pi_j}$	Infinitesimal generator of $(\hat{X}_i^{\Pi_i,\Pi_j}(t), Y(t))$.
μ_i	Mean of the distribution Π_i.
σ_i^2	Variance of the distribution Π_i.
$Q_{\Pi_i}(p)$	Left-quantile function of the distribution Π_i.
h_i	Distortion function of Choquet regularizer for Agent i.
Φ_{h_i}	Choquet regularizer for Agent i.
V_i, g_i	Value functions of Agent i.
Θ	Parameters of an approximate value function.
$V_i^{\Theta}, g_i^{\Theta}$	Parametrized approximations of value functions V_i and g_i.
Ψ	Parameters of an approximate policy.
Π_i^{Ψ}	Parametrized approximation of Π_i.

Acknowledgements. The research of Junyi Guo was supported by the National Natural Science Foundation of China (No. 12271274). The research of Xia Han was supported by the National Natural Science Foundation of China (Grant Nos. 12301604, 12371471, and 12471449) and the Fundamental Research Funds for the Central Universities, Nankai University (Grant No. 63231138).

Competing interests. The authors declare none.

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

References

Abel, A.B. (1990) Asset prices under habit formation and catching up with the Joneses. *The American Economic Review*, **80**, 38–42.

Basak, S. and Chabakauri, G. (2010) Dynamic mean-variance asset allocation. *The Review of Financial Studies*, **23**(8), 2970–3016.

Bensoussan, A., Siu, C, Yam, S. and Yang, H. (2014) A class of non-zero-sum stochastic differential investment and reinsurance games. *Automatica*, **50**(8), 2025–2037.

Björk, T., Khapko, M. and Murgoci, A. (2017) On time-inconsistent stochastic control in continuous time. *Finance and Stochastics*, **21**, 331–360.

Björk, T. and Murgoci, A. (2010) *A general theory of Markovian time inconsistent stochastic control problems*, Stockholm School of Economics, working paper.

Björk, T. and Murgoci, A. (2014) A theory of Markovian time-inconsistent stochastic control in discrete time. *Finance and Stochastics*, **18**, 545–592.

Björk, T., Murgoci, A. and Zhou, X. (2014) Mean-variance portfolio optimization with state-dependent risk aversion. *Mathematical Finance*, **24**(1), 1–24.

Browne, S. (2000) Stochastic differential portfolio games. *Journal of Applied Probability*, **37**(1), 126–147

Chen, L. and Shen, Y. (2019) Stochastic Stackelberg differential reinsurance games under time-inconsistent mean-variance framework. *Insurance: Mathematics and Economics*, **88**, 120–137.

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

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

DeMarzo, P.M., Kaniel, R. and Kremer, I. (2008) Relative wealth concerns and financial bubbles. *The Review of Financial Studies*, **21**(1), 19–50.

Deng, C., Zeng, X. and Zhu, H. (2018) Non-zero-sum stochastic differential reinsurance and investment games with default risk. *European Journal of Operational Research*, **264**(3), 1144–1158.

Ekeland, I. and Pirvu, T.A. (2008) Investment and consumption without commitment. *Mathematics and Financial Economics*, **2**(1), 57–86.

Espinosa, G. and Touzi, N. (2015) Optimal investment under relative performance concerns. *Mathematical Finance*, **25**(2), 221–257.

Foerster, J., Nardelli, N., Farquhar, G., Afouras, T., Torr, P.H., Kohli, P. and Whiteson, S. (2017) Stabilising experience replay for deep multi-agent reinforcement learning. *International Conference on Machine Learning*, pp. 1146–1155.

Föllmer, H. and Schied, A. (2011) *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter.

Gali, J. (1994) Keeping up with the Joneses: Consumption externalities, portfolio choice, and asset prices. *Journal of Money, Credit and Banking*, **26**(1), 1–8.

Gilboa, I. and Schmeidler, D. (1989) Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, **18**(2), 141–153.

Guo, J., Han, X. and Wang, H. (2025) Exploratory mean-variance portfolio selection with Choquet regularizers. *Quantitative Finance*, 1–21.

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

Hu, D. and Wang, H. (2018) Time-consistent investment and reinsurance under relative performance concerns. *Communications in Statistics-Theory and Methods*, **47**(7), 1693–1717.

Hu, T. and Chen, O. (2020) On a family of coherent measures of variability. *Insurance: Mathematics and Economics*, **95**, 173–182.

Isaacs, R. (1965) *Differential Games*. New York: Wiley.

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

Jia, Y. and Zhou, X.Y. (2022b) Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, **23**(275), 1–50.

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

Kim, T.S. and Omberg, E. (1996) Dynamic nonmyopic portfolio behavior. *The Review of Financial Studies*, **9**(1), 141–161.

Li, D. and Ng, W.L. (2000) Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, **10**, 287–406.

Li, D. and Young, V.R. (2021) Bowley solution of a mean-variance game in insurance. *Insurance: Mathematics and Economics*, **98**, 35–43.

Littman, M.L. (1994) Markov games as a framework for multi-agent reinforcement learning. In Machine Learning Proceedings 1994, pp. 157–163.

Littman, M.L. (2001) Friend-or-foe Q-learning in general-sum games. *International Conference on Machine Learning*, pp. 322–328.

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

Liu, J. (2001) *Dynamic portfolio choice and risk aversion*, *working paper*, UCLA.

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

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

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press

Pontryagin, L.S. (1967) Linear differential games. I, II, in: *Doklady Akademii Nauk. Russian Academy of Sciences*, **175**, 764–766.

Quiggin, J. (1982) A theory of anticipated utility. *Journal of Economic Behavior and Organization*, **3**(4), 323–343.

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

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, **57**(3), 571–587.

Siu, C.C., Yam, S.C.P., Yang, H. and Zhao, H. (2016) A class of nonzero-sum investment and reinsurance games subject to systematic risks. *Scandinavian Actuarial Journal*, **2017**(8), 670–707.

Sun, Z. and Jia, G. (2023) Reinforcement learning for exploratory linear-quadratic two-person zero-sum stochastic differential games. *Applied Mathematics and Computation*, **442**, 127763.

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

Wachter, J.A. (2002) Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets. *Journal of Financial and Quantitative Analysis*, **37**(1), 63–91.

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

Wang, H., Zariphopoulou, T. and Zhou, X.Y. (2020a) Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, **21**(1), 8145–8178.

Wang, Q., Wang, R. and Wei, Y. (2020b). Distortion risk metrics on general spaces. *ASTIN Bulletin*, **50**(4), 827–851.

Wang, R., Wei, Y. and Willmot, G.E. (2020c) Characterization, robustness and aggregation of signed Choquet integrals. *Mathematics of Operations Research*, **45**(3), 993–1015.

Wang, N., Zhang, N., Jin, Z. and Qian, L. (2019) Robust non-zero-sum investment and reinsurance game with default risk. *Insurance: Mathematics and Economics*, **84**, 115–132.

Wang, N., Zhang, N., Jin, Z. and Qian, L. (2021). Reinsurance-investment game between two mean-variance insurers under model uncertainty. *Journal of Computational and Applied Mathematics*, **382**, 113095.

Yaari, M. E. (1987) The dual theory of choice under risk. *Econometrica*, **55**(1), 95–115.

Yang, Y. and Wang, J. (2020) An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv:2011.00583.

Zeng, Y., Li, D. and Gu, A. (2016) Robust equilibrium reinsurance-investment strategy for a mean-variance insurer in a model with jumps. *Insurance: Mathematics and Economics*, **66**, 138–152.

Zhang, K., Yang, Z. and Basar, T. (2021) Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*, pp. 321–384.

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

Zhu, J., Guan, G. and Li, S. (2020) Time-consistent non-zero-sum stochastic differential reinsurance and investment game under default and volatility risks. *Journal of Computational and Applied Mathematics*, **374**, 112737.

https://doi.org/10.1017/asb.2025.10080 Published online by Cambridge University Press