NUS RMI Working Paper Series – No. 2021-02

Learning Equilibrium Mean-Variance Strategy

Min DAI, Yuchao DONG, and Yanwei JIA

March 2021

NUS Risk Management Institute
21 HENG MUI KENG TERRACE, #04-03 I3 BUILDING, SINGAPORE 119613

www.rmi.nus.edu.sg/research/rmi-working-paper-series

Learning Equilibrium Mean-Variance Strategy

Min Dai
Department of Mathematics, RMI, and NUSRI, National University of Singapore, Singapore, Singapore 119076.
mindai@nus.edu.sg

Yuchao Dong
Institute of Operations Research and Analytics, National University of Singapore, Singapore, Singapore 119076.
orady@nus.edu.sg

Yanwei Jia
Department of Mathematics, National University of Singapore, Singapore, Singapore 119076. jia_yanwei@u.nus.edu

We study a dynamic mean-variance portfolio optimization problem under the reinforcement learning framework, where an entropy regularizer is introduced to induce exploration. Due to the time-inconsistency involved in a mean-variance criterion, we aim to learn an equilibrium strategy. Under an incomplete market setting, we obtain a semi-analytical, exploratory, equilibrium mean-variance strategy that turns out to follow a Gaussian distribution. We then focus on a Gaussian mean return model and propose an algorithm to find the equilibrium strategy using the reinforcement learning technique. Thanks to a thoroughly designed policy iteration procedure in our algorithm, we can prove our algorithm's convergence under mild conditions, despite that dynamic programming principle and the usual policy improvement theorem fail to hold for an equilibrium solution. Numerical experiments are given to demonstrate our algorithm.

Key words: asset allocation; reinforcement learning; equilibrium mean variance analysis; entropy regularized exploration-exploitation

History: This paper was first submitted on May 12, 2020.

1. Introduction

Mean-variance portfolio optimization, founded by Markowitz (1952), marks the beginning of modern finance and becomes a key component of data-based investment. However, almost all practical applications of mean-variance analysis have been restricted to static or myopic investment (cf. Black and Litterman (1991)), partially due to two barriers. First, a mean-variance analysis has an inherent time-inconsistency issue such that most of the dynamic mean-variance strategies are

Electronic copy available at: https://ssrn.com/abstract=3770818

either time-inconsistent or not conforming with conventional investment wisdom (Dai et al. 2020). Second, a dynamic mean-variance model is sensitive to parameter values, but calibrating a dynamic model is notoriously difficult (cf. Merton (1980), Luenberger (1998)).

Driven by recent trends in financial innovation and the needs in the big data era, we aim to design an efficient algorithm to learn a dynamic, time-consistent mean-variance strategy by virtue of mass financial data. More specifically, we will combine the concept of equilibrium mean-variance solution and the reinforcement learning (RL, for short) technique to overcome the two barriers aforementioned.

This work is motivated by two recent developments in this area. The first is made by Dai et al. (2020) who propose a dynamic, equilibrium mean-variance criterion for portfolio's log-returns (hereafter log-MV criterion, for short).[1] The criterion leads to an analytical, time-consistent strategy that complies with conventional investment wisdom even in some incomplete markets. Moreover, one can easily elicit investors' risk aversion preference associated with the criterion, which makes promising a model-based robo-advising system. We will follow Dai et al. (2020) to consider the log-MV criterion and seek an equilibrium (time-consistent) strategy. Different from Dai et al. (2020) with given model parameters, we assume unknown model parameters and aim to learn the strategy in terms of RL.

The second development is made by Wang et al. (2019) who propose a general continuous-time exploratory stochastic control framework with RL. They introduce entropy to characterize exploration in a control formulation and develop an RL algorithm to learn optimal strategy. Wang and Zhou (2020) apply this framework for a pre-committed mean-variance strategy and find a significant improvement over traditional MLE-based methods. In this paper, we will borrow their idea to consider an equilibrium mean-variance solution. It is worthwhile to point out that finding an equilibrium solution is not standard in the RL literature, as there is neither dynamic programming principle nor an objective functional to optimize.

Our major contributions can be summarized as follows.

Electronic copy available at: https://ssrn.com/abstract=3770818

(i) We extend the exploratory stochastic control framework in Wang et al. (2019) to an incomplete market, where the asset return is correlated with a stochastic market state. The extension is not restricted to mean-variance portfolio optimization problems. It is interesting to notice that to derive an "exploratory" market dynamics in the incomplete market, a new Brownian motion that is independent of the market is introduced into the market dynamics associated with randomized control action. Our formulation indicates that the new Brownian motion is to model the noise caused by exploration. For a complete market, we are able to recover the single Brownian motion involved in Wang et al. (2019) for their exploratory market dynamics, but we highlight that the Brownian motion is essentially different from the original Brownian motion driving the market. For details, see Section 2.2.

(ii) We obtain a semi-analytical solution for the equilibrium log-MV problem under the exploratory framework in the incomplete market. Interestingly, we find that the exploratory equilibrium strategy also follows a Gaussian distribution whose mean coincides with the (non-exploratory) strategy studied in Dai et al. (2020). Moreover, its variance is proportional to the exploration-exploitation parameter, the reciprocal of one's risk preference, and the reciprocal of instantaneous variance of the stock price. This suggests that if the proposed exploratory problem is solved effectively, then one may learn the true equilibrium solution by taking the expectation of the exploratory strategies. For details, see Section 3.

(iii) For algorithm design and numerical implement, we take for example the Gaussian mean return model, for which we obtain a closed-form, exploratory, equilibrium mean-variance solution. It should be pointed out that different from the traditional policy iteration procedure arising from a dynamic optimization problem (e.g. Wang and Zhou (2020)), a policy iteration for an equilibrium solution usually does not possess the nice property of policy improvement. We propose a novel policy iteration procedure with a carefully selected initial guess and prove its convergence to the desired equilibrium solution under mild conditions. Our proof provides a new perspective to analyze RL algorithms by generalizing the iteration of parameters to

Electronic copy available at: https://ssrn.com/abstract=3770818

the iteration of functions that are embedded into a suitable function space. For details, see Section 4. Based on the policy iteration procedure, we develop an RL algorithm to learn the equilibrium strategy. To the best of our knowledge, this paper is the first one to study equilibrium strategy in terms of RL. Numerical results on simulated data set demonstrate the efficiency and robustness of our algorithm; see Section 6.

Related Literature

The original mean-variance portfolio optimization problem formulated by Markowitz (1952) is for a single period. It is challenging to extend the mean-variance analysis to a dynamic setting due to the inherent time-inconsistency issue, i.e., a mean-variance strategy that is optimal today may not be optimal tomorrow.[2] A naive way of handling the time-inconsistency in a multi-period setting is to optimize a myopic mean-variance criterion at each step, ignoring its dynamic nature and rolling up until maturity (e.g. Aït-Sahali and Brandt (2001) and Campbell and Viceira (2002)). Unfortunately, such a myopic strategy turns out to be significantly sub-optimal in many incomplete markets (e.g. Kim and Omberg (1996), Brandt (1999), and Campbell and Viceira (1999)). A more sophisticated way is to seek the so-called pre-committed mean-variance strategy that is optimal only at initial time, disregarding subsequent sub-optimality and time-inconsistency (e.g. Li and Ng (2000) and Zhou and Li (2000)). However, a pre-committed strategy, while applied to a robo-advising system, may confuse the general public due to the time-inconsistency.

Basak and Chabakauri (2010) provide an alternative way to attack the time-inconsistency involved in a mean-variance criterion: though the problem itself is time-inconsistent, one may seek a time-consistent strategy. The idea is later extended by Björk et al. (2017) to provide a general framework for handling time-inconsistency, which leads to the so-called equilibrium strategy that can be regarded as a subgame perfect Nash equilibrium in dynamic games. In this paper, we focus on the equilibrium strategy as we believe that time-consistency is a fundamental requirement for rational decision making. In particular, we adopt the log-MV criterion proposed by Dai et al. (2020) because equilibrium strategies associated with other existing mean-variance criteria may

Electronic copy available at: https://ssrn.com/abstract=3770818

conflict with some of conventional investment wisdom (cf. Basak and Chabakauri (2010), Björk et al. (2014), and Dai et al. (2020)). The major difference between this paper and Dai et al. (2020) lies in that we work with an exploratory framework and aim at developing an RL algorithm to find equilibrium solution without knowing market parameters.

This paper is related to a strand of literature on RL, which is about how a software agent interacts with environment to maximize her payoff. The agent's actions serve as a mean to both explore (learn) and exploit (optimize). In some sense, RL combines parameter estimation and optimization together. Nowadays, RL has become one of the most active and fast developing areas of machine learning, due to its success in playing go (Silver et al. 2016, 2017), video games (Mnih et al. 2015), controlling robotics (Deisenroth et al. 2013), designing autonomous driving (Levine et al. 2016) and so on. Such huge success also draws attentions of both researchers and practitioners in the financial industry, such as Nevmyvaka et al. (2006) and Hendricks and Wilcox (2014) on trade execution, Moody et al. (1998) and Moody and Saffell (2001) on algorithmic trading, and Guéant and Manziuk (2019) on market making. In particular, there are many attempts to apply RL in dynamic portfolio optimization; see, e.g. Neuneier (1996), Gao and Chan (2000), Jin and El-Saawy (2016), and Ritter (2017), but these studies are all under the utility maximization framework. In contrast, we consider mean-variance portfolio optimization that is intuitively more appealing.

Extant literature on mean-variance portfolio optimization with RL has been restricted to the pre-committed strategy; see, e.g. Sobel (1982), Sato and Kobayashi (2000), Sato et al. (2001), Tamar and Mannor (2013), Prashanth and Ghavamzadeh (2013, 2016), and Wang and Zhou (2020). Compared with the literature, we aim to look for an equilibrium mean-variance strategy with RL. Moreover, we allow the market to be incomplete with an observable stochastic factor.

Most of existing algorithms, like the ϵ-greedy algorithm and its variants, treat exploration separately as an exogenous ad-hoc choice, rather than include exploration as part of optimization objective (e.g. Tokic (2010) and Sutton and Barto (2011)). A discrete time entropy-regularized RL framework is proposed in Ziebart et al. (2008) and Todorov (2007), where exploration is incorporated into the optimization objective (see also Azar et al. (2011), Fox et al. (2015), and Nachum

Electronic copy available at: https://ssrn.com/abstract=3770818

et al. (2017)). Wang et al. (2019) propose a continuous-time entropy-regularized RL framework, where the policy is extended to a probability-measure-valued process, and they show that the optimal exploratory control policies must follow a Gaussian distribution when the objective reward is a quadratic function. As an application, Wang and Zhou (2020) show that under the Black-Scholes market, the exploratory, pre-committed mean-variance strategy follows a Gaussian distribution. In this paper, following Wang et al. (2019), we incorporate an entropy-regularizer into a continuous-time mean-variance criterion for portfolio's log return proposed by Dai et al. (2020) and learn an equilibrium mean-variance strategy that proves to follow a Gaussian distribution even under an incomplete market setting. Moreover, we find that as with Wang et al. (2019) for the Black-Scholes market, our exploratory formulation under an incomplete market is still linked to the relaxed stochastic control theory (cf. Fleming and Nisio (1984) and Zhou (1992)).

Most RL algorithms are based on the dynamic programming principle, e.g., the well-known Q-learning and its variants (cf. Watkins and Dayan (1992), Van Hasselt et al. (2016), and Doya (2000)). The dynamic programming principle, however, cannot be applied directly to mean-variance portfolio optimization. Zhou and Li (2000) employ an embedding technique such that the dynamic programming principle can still be applied to find a pre-committed strategy. Sato et al. (2001) introduce a TD-learning algorithm to estimate the variance and a gradient-based RL algorithm on the mean-variance problem. Using a linear function approximation, Tamar and Mannor (2013) present an actor-critic algorithm and prove its convergence to a locally optimal point. Their method is further developed by Prashanth and Ghavamzadeh (2013, 2016) and Wang and Zhou (2020). We combine the idea of TD-error and the actor-critic approach to approximate and evaluate policies and value functions simultaneously. As in Wang and Zhou (2020), we also propose a simple parametrization method that mimics the theoretical solution, without involving neutral networks. More importantly, we present an algorithm for the equilibrium solution for time-inconsistent problems.

The rest of the paper is organized as follows. In Section 2, we present the market setup and introduce an exploratory, entropy-regularized, mean-variance problem as well as the definition

Electronic copy available at: https://ssrn.com/abstract=3770818

of the associated equilibrium solution. A semi-analytical, exploratory, equilibrium mean-variance strategy is given in Section 3. In Section 4, we focus on the Gaussian mean return model and present a policy iteration procedure as well as its convergence analysis. An RL algorithm based on the convergence analysis is proposed in Section 5. Numerical results are given to demonstrate the algorithm in Section 6. We conclude in Section 7. All technical proofs and additional results are relegated to E-Companion.

2. Model Setup

2.1. Market Environment

Assume two assets available for investment: a riskless asset (bond) with interest rate r and a risky asset (stock). The stock price is governed by:

$$\frac{\mathrm{d}S_t}{S_t} = \mu_t \mathrm{d}t + \sigma_t \mathrm{d}B_t, \tag{1}$$

where B_t is a scalar-valued Brownian motion, and μ_t and σ_t depend on a stochastic market state X_t. We further assume that X_t is a diffusion process satisfying

$$\mathrm{d}X_t = m_t \mathrm{d}t + \nu_t[\rho \mathrm{d}B_t + \sqrt{1-\rho^2}\mathrm{d}\tilde{B}_t], \tag{2}$$

where ρ is a constant and $\tilde{B}_t$ is another (scalar-valued) Brownian motion that is independent of B_t. We assume that $\mu_t \equiv \mu(t, X_t)$, $\sigma_t \equiv \sigma(t, X_t)$, $m_t \equiv m(t, X_t)$, and $\nu_t \equiv \nu(t, X_t)$ are all deterministic functions, which, however, we do not know.

Note that the above market model, similar to that in Dai et al. (2020), is very general and covers many popular models as our special cases. For example, it covers the Gaussian mean return model and the stochastic volatility model discussed in Wachter (2002) and Liu (2007), respectively.

A self-financing wealth process W_t can be described as

$$\frac{dW_t}{W_t} = [r + (\mu_t - r)u_t]dt + \sigma_t u_t dB_t,$$

where u_t, representing the fraction of total wealth invested in stock at time t, is a scalar-valued adapted process and can be regarded as a strategy. For later use, we introduce the portfolio's log return process $R_t = \log W_t$, which satisfies:

$$\mathrm{d}R_t = \left[r_t + (\mu_t - r_t)u_t - \frac{1}{2}\sigma_t^2 u_t^2\right]\mathrm{d}t + \sigma_t u_t \mathrm{d}B_t. \tag{3}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

2.2. An Exploratory Version of the Portfolio Return Process

Let T be the investment horizon. Recall the dynamic mean-variance criterion proposed in Dai et al. (2020), which is a trade-off between mean and variance of the log return of the portfolio R_T, namely,

$$\mathbb{E}_t\left[R_T\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_T\right].$$

Here $\mathbb{E}_t[\cdot]$ and $\mathrm{Var}_t[\cdot]$ stand for the conditional expectation and variance at time t, respectively, and γ measures the trade-off between mean and variance. This criterion can yield an analytical equilibrium strategy that conforms to common investment wisdom (cf. Dai et al. (2020) for more discussion). We will incorporate learning (i.e. exploration) into the mean-variance formulation.

Consider an investor who aims to seek a dynamic mean-variance equilibrium strategy through RL. The investor tries to learn the "best" strategy based on learning by doing, through the interactions with the environment. Due to information incompleteness, the investor has to make a balance between exploration, through which she gathers more information that might lead to better decisions, and exploitation, through which she makes the best decision with current information. One way to do exploration is adding a small noise to each action (namely, strategy) taken. Following Wang et al. (2019), we randomize the strategy process u_t and result in a distributional strategy process whose density function is as given by $\{\pi_t,\ 0 \le t \le T\}$.

As our market set up involves the market state X_t, the "exploration" version of the portfolio return process associated with distributional strategy process is different from that in Wang et al. (2019). To give an intuition about what the dynamics should look like, let us examine the discrete time case at time t:

$$\Delta R_t = \left[r + (\mu - r)u - \frac{1}{2}\sigma^2 u^2\right]\Delta t + \sigma u \Delta B_t.$$

Let u be sampled from an independent distribution π, and consider the following increment:

$$\tilde{\Delta} := \left[r + (\mu - r)\int_{\mathbb{R}} u\pi(du) - \frac{1}{2}\sigma^2\int_{\mathbb{R}} u^2\pi(du)\right]\Delta t + \sigma\int_{\mathbb{R}} u\pi(du)\Delta B_t + \sqrt{\int_{\mathbb{R}} u^2\pi(du) - (\int_{\mathbb{R}} u\pi(du))^2}\Delta\bar{B}_t$$

Electronic copy available at: https://ssrn.com/abstract=3770818

with $\bar{B}_t$ being another Brownian motion. It follows

$$\begin{aligned}\mathbb{E}\left[\tilde{\Delta}\right] &= \left[r+(\mu-r)\int_{\mathbb{R}} u\pi(du)-\frac{1}{2}\sigma^2\int_{\mathbb{R}} u^2\pi(du)\right]\Delta t,\\ \mathrm{Var}\left[\tilde{\Delta}\right] &= \sigma^2\int_{\mathbb{R}} u^2\pi(du)\Delta t+o(\Delta t),\quad \mathrm{Cov}\left[\tilde{\Delta},\Delta X_t\right]=\rho\nu\sigma\int_{\mathbb{R}} u\pi(du)\Delta t+o(\Delta t).\end{aligned}$$

It is easy to see that $\tilde{\Delta}$ approximates ΔR_t on first and second moments.

Motivated by the above observation, we replace (3) by the following process that is associated with randomized strategy characterized by π and will be used in the exploratory mean-variance formulation:

$$\begin{aligned}\mathrm{d}R_t^\pi = &\left[r+(\mu_t-r)\int_{\mathbb{R}} u\pi_t(\mathrm{d}u)-\frac{1}{2}\sigma_t^2\int_{\mathbb{R}} u^2\pi_t(\mathrm{d}u)\right]\mathrm{d}t\\ &+\sigma_t\left[\int_{\mathbb{R}} u\pi_t(\mathrm{d}u)\mathrm{d}B_t+\sqrt{\int_{\mathbb{R}} u^2\pi_t(\mathrm{d}u)-(\int_{\mathbb{R}} u\pi_t(\mathrm{d}u))^2}\mathrm{d}\bar{B}_t\right],\end{aligned}\tag{4}$$

where $\bar{B}_t$ is another Brownian motion that is mutually independent of B_t and $\tilde{B}_t$.

Equation (4) characterizes the impact of the strategy on the portfolio return process. It is worthwhile pointing out that in a complete market where μ_t and σ_t do not depend on the market state X_t, we can merge two Brownian motions of Equation (4) into one Brownian motion and thus recover the formulation as given in Wang et al. (2019).[3] In an incomplete market, it is interesting to note that (4) involves a new Brownian motion $\bar{B}_t$. Intuitively speaking, B_t and $\tilde{B}_t$ in (4) and (2) are used to model the market noises, while $\bar{B}_t$ is introduced to model the noise caused by exploration and can be regarded as a "random number generator" that the investor uses to generate a random strategy. The coefficient of $\mathrm{d}\bar{B}_t$ term reflects the variance of π_t, measuring how much additional noise is introduced into the system. Later we will see that consistent with the observation of Wang et al. (2019), our new dynamic system also falls within the relaxed control framework in Fleming and Nisio (1984) and Zhou (1992), where control policies are extended to probability distributions.

2.3. Entropy-regularized Mean Variance Problem

We now follow Wang et al. (2019) to incorporate an entropy regularizer into the mean-variance criterion over R_t^π and get the following reward functional:

$$J(t,R_t^\pi,X_t;\pi):=\mathbb{E}_t\left[R_T^\pi+\lambda\int_t^T H(\pi_s)ds\right]-\frac{\gamma}{2}\mathrm{Var}_t\left[R_T^\pi\right].\tag{5}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

where R_t^π and X_t follow (4) and (2), respectively, λ represents the exploration weight that will be tuned to achieve a nearly best trade-off between exploration and exploitation,[4] and H is the entropy of the strategy distribution as defined below:

$$H(\pi) = \begin{cases} -\int_{\mathbb{R}} \pi(u) \log \pi(u) du, \text{ if } \pi(du) = \pi(u)du, \\ -\infty, \text{ otherwise.} \end{cases}$$

It should be emphasized that time inconsistency is an inherent issue of a mean-variance problem. We will follow Björk et al. (2017) to consider an equilibrium mean-variance solution associated with the entropy-regularized mean-variance criterion (5). Before introducing the definition of an equilibrium solution, let us first define admissible feedback control $\pi_s = \pi(s, R_s, X_s)$ as we are interested in feedback strategy.

DEFINITION 1. $\pi = \{\pi_s, t \le s \le T\}$ is called an **admissible feedback control**, if

(i) for each $t \le s \le T$, $\pi_s \in \mathcal{P}(\mathbb{R})$ a.s., where $\mathcal{P}(\mathbb{R})$ stands for all probability measures on the real numbers;

(ii) $\pi_s = \pi(s, R_s, X_s)$, where $\pi(\cdot, \cdot, \cdot)$ is a deterministic mapping from $[t, T] \times \mathbb{R} \times \mathbb{R}$ to $\mathcal{P}(\mathbb{R})$;

(iii) $\mathbb{E}_t \left[\int_t^T \int_{\mathbb{R}} |\sigma_s u|^2 \pi_s(du) ds \right] + \mathbb{E}_t \left[\int_t^T \int_{\mathbb{R}} |\mu_s u| \pi(du) ds \right] < \infty$.

The collection of all the admissible controls in the feedback form at time t is denoted as Π_t.

We are ready to define the equilibrium solution as follows:

DEFINITION 2. An admissible control π^* is called an equilibrium policy, if, at any time t, for any perturbation control $\pi^{h,v}$ defined by

$$\pi_\tau^{h,v} = \begin{cases} v, \text{ for } t \le \tau \le t + h, \\ \pi_\tau^*, \text{ for } t + h \le \tau \le T. \end{cases}$$

with any $h \in \mathbb{R}^+$ and $v \in \mathcal{P}(\mathbb{R})$, the entropy-regularized mean-variance functional is locally better off, namely

$$\liminf_{h \to 0^+} \frac{J(t, R_t, X_t; \pi^*) - J(t, R_t, X_t; \pi^{h,v})}{h} \ge 0.$$

Furthermore, for an equilibrium control π^*, the equilibrium value function V^* is defined as

$$V^*(t, R_t, X_t) := J(t, R_t, X_t; \pi^*).$$

Electronic copy available at: https://ssrn.com/abstract=3770818

An equilibrium policy that is optimal locally at any time given that the policy will be followed in the future. It is consistent with the subgame perfect Nash equilibrium in the dynamic games in economics.

3. Equilibrium Strategy

In this section, we provide a characterization of equilibrium solution for the general incomplete case. The complete market case is much simpler and is relegated to E-Companion EC.2.

The following theorem shows that a semi-analytical equilibrium policy exists in general under mild conditions.

THEOREM 1. *Assume that* $\mathbb{E}\left[\int_0^T \theta_s^2 ds\right]$ *and* $\mathbb{E}\left[e^{\frac{\gamma^2}{1+\gamma^2}(\int_0^T (r_t+\theta^2/2)dt+\int_0^T \theta_t/\rho dB_t^X}\right]<\infty$, *where* $\theta_t = \theta(t,X_t)=\frac{\mu(t,X_t)-r}{\sigma(t,X_t)}$ *and* $dB_t^X=\rho dB_t+\sqrt{1-\rho^2}d\tilde{B}_t$. *Then, we have the following results:*

(i) An equilibrium policy is given by

$$\pi^*(t,X)\sim\mathcal{N}(\frac{\mu_t-r_t}{(1+\gamma)\sigma_t^2}-\frac{\rho\gamma Z_t}{(1+\gamma)\sigma_t},\frac{\lambda}{(1+\gamma)\sigma_t^2}), \tag{6}$$

where Z_t *is uniquely determined by the following backward stochastic differential equation (BSDE):*

$$dY_t=-f(t,X_t,Z_t)dt+Z_t dB_t^X,\quad Y_T=0 \tag{7}$$

with $f(t,X,Z)=r-\frac{\lambda}{2(1+\gamma)}+\frac{1}{2}\theta^2(t,X)-\frac{\gamma^2(\theta(t,X)+\rho Z)^2}{2(1+\gamma)^2}$.

(ii) There exists a deterministic function $h(\cdot,\cdot)$ *such that* $Y_t=h(t,X_t)$. *Moreover, if* $h\in C^{1,2}$, *then* h *solves the following partial differential equation (PDE):*

$$\partial_t h+m\partial_X h+\frac{1}{2}\nu^2\partial_{XX}h+f(t,X,\nu\partial_X h)=0,\quad h(T,X)=0, \tag{8}$$

and $Z_t=\nu\partial_x h(t,X_t)$.

(iii) Under the equilibrium policy (6), *we have*

$$\begin{aligned} &V^*(t,R,X)\equiv J(t,R,X;\pi^*)=U^*(t,X)+R;\\ &g^*(t,R,X)\equiv\mathbb{E}_t[R_T^{\pi^*}|R_t^{\pi^*}=R,X_t=X]=h^*(t,X)+R. \end{aligned} \tag{9}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

for some function h^* and U^* with $h^*(T,X)=U^*(T,X)=0$. Moreover, if $h^*, U^* \in C^{1,2}$, then h^* satisfies (8) *and U^* satisfies*

$$\partial_t U + m\partial_X U + \frac{1}{2}\nu^2\partial_{XX}U + r + \frac{(\mu - r - \rho\gamma\nu\sigma\partial_X h^*)^2}{2(1+\gamma)\sigma^2} + \frac{1}{2}\log\frac{2\pi\lambda}{(1+\gamma)\sigma^2} - \frac{\gamma}{2}\nu^2|\partial_X h^*|^2 = 0. \tag{10}$$

The above theorem indicates that under mild regularity conditions, one can construct an equilibrium policy by solving a BSDE. Note that even in an incomplete market, the distribution of the exploratory equilibrium strategies is still Gaussian. Moreover, the mean of the exploratory strategies turns out to be independent of the exploration weight λ. This suggests a separation of the exploration and exploitation, which is also observed in Wang et al. (2019) and Wang and Zhou (2020).

Equations (7) and (8) are almost the same as the corresponding equations in Dai et al. (2020), except that function f in the former involves the exploration weight λ. On the other hand, the variance of exploratory strategies is proportional to the exploration weight, as well as the reciprocal of the risk aversion and instantaneous variance of stock return. This implies that as the investor becomes more risk averse or the market gets more volatile, the equilibrium policy will get more concentrated and the investor is less willing to explore environment.

Unlike the pre-committed strategy discussed in Wang and Zhou (2020), the variance of our equilibrium solution does not necessarily decay in time. For instance, if the stock volatility is a constant, then the variance of our equilibrium strategy remains a constant. This result is partially due to two reasons. First, given the requirement of a subgame perfect equilibrium, the "future self" and the "current self" will not coordinate on exploration due to the lack of self-commitment, which causes inefficiency.[5] Second, there is no discount term in the accumulative reward for exploration in our objective functional (5), which indicates that exploration at any time would be rewarded equally. Therefore, it looks as if the investor would choose a myopic exploration at any time, according to the backward induction procedure. To incorporate increasing exploration experience

Electronic copy available at: https://ssrn.com/abstract=3770818

into our model, one could simply add a time-decaying exploration parameter in the running reward. Such a design is discussed in E-Companion EC.1.

The BSDE approach used here for studying equilibrium policy is first introduced in Dai et al. (2020). Compared with the PDE approach in Björk et al. (2017), the BSDE approach usually requires less regularity of solution to a system of Hamilton–Jacobi–Bellman (HJB) equations. The PDE approach for a verification theorem will be discussed in E-Companion EC.5, where Equations (8) and (10) also appear in the system of HJB equations. Therefore, the BSDE approach and the PDE approach share the same PDE system if a classical solution is admitted.

As an application of Theorem 1, we consider the time varying Gaussian mean return model, namely

$$\mu(t,X)=r+\sigma X, \sigma(t,X)=\sigma, m(t,X)=\iota(\bar{X}-X) \text{ and } \nu(t,X)=\nu,$$

where r, σ, ι, $\bar{X}$, and ν are all positive constants. The following Proposition shows that a closed-form equilibrium solution is available for this model.

PROPOSITION 1. *For the Gaussian mean return model, an equilibrium strategy is*

$$\pi^*(t,X)\sim\mathcal{N}\Big(\frac{X}{(1+\gamma)\sigma}-\frac{\gamma\rho\nu}{(1+\gamma)\sigma}(a_2^*(t)X+a_1^*(t)),\frac{\lambda}{(1+\gamma)\sigma^2}\Big) \tag{11}$$

with

$$\begin{cases} a_2^*(t)=\dfrac{(1+2\gamma)}{(1+\gamma)^2}\dfrac{e^{2C_1(T-t)}-1}{(C_1+C_2)(e^{2C_1(T-t)}-1)+2C_1} \\ a_1^*(t)=\dfrac{\iota\bar{X}(1+2\gamma)}{(1+\gamma)^2}\dfrac{(e^{C_1(T-t)}-1)^2}{C_1[(C_1+C_2)(e^{2C_1(T-t)}-1)+2C_1]}, \end{cases} \tag{12}$$

where $C_1=\frac{1}{\gamma+1}\left[\gamma^2(\iota+\rho\nu)^2+\iota^2(2\gamma+1)\right]^{\frac{1}{2}}$ *and* $C_2=\iota+\frac{\gamma^2\rho\nu}{(1+\gamma)^2}$.

Moreover, the associated V^* *and* g^* *have the following form:*

$$\begin{aligned} V^*(t,R,X)&=R+\frac{1}{2}b_2^*(t)X^2+b_1^*(t)X+b_0^*(t),\\ g^*(t,R,X)&=R+\frac{1}{2}a_2^*(t)X^2+a_1^*(t)X+a_0^*(t), \end{aligned} \tag{13}$$

where $a_2^*(t)$ *and* $a_1^*(t)$ *are as given by (12), and* $a_0^*(t)$, $b_1^*(t)$, *and* $b_2^*(t)$ *are given in E-Companion* (EC.13) *and* (EC.14).

Electronic copy available at: https://ssrn.com/abstract=3770818

It can be observed that for the Gaussian mean return model, the mean of the exploratory equilibrium strategy coincides with the classical equilibrium strategy studied in Dai et al. (2020). It means that if an investor follows the mean strategy to invest, then she would obtain the same strategy as one who knows market parameters. Moreover, the variance of exploration in learning indeed remains at a constant level. Hence, when designing an algorithm, we shall simply look for a constant instead of a complex function as the exploration variance. This may greatly simplify the algorithm.

An application to the stochastic volatility model is presented in E-Companion EC.4.

4. Numerical Methods

Let us first describe the basic idea about how to use the RL technique to find equilibrium policy.

4.1. General Discussions

In the standard RL algorithm (Sutton and Barto 2011), a learning procedure usually consists of two iterative steps:

(i) Given a policy π, compute the associated value function V^π;[6]

(ii) Update the previous policy π to a new one $\tilde{\pi}$ according to the obtained value function V^π.

We want to employ a similar procedure to design a numerical algorithm. Therefore, it is important to understand two questions: (1) what is the value function V^π associated with a feasible policy? and (2) what is the criterion to update the policy?

Let $\pi_t = \pi(t, R_t, X_t)$ be an admissible policy. The value function associated with the policy π is defined as $V^\pi(t, R_t, X_t) = J(t, R_t^\pi, X_t; \pi)$, and an auxiliary value function is defined as

$$g^\pi(t, R, X) := \mathbb{E}_t \left[R_T^\pi | R_t^\pi = R, X_t = X \right],$$

where R_t^π is the portfolio return process associated with the control policy π.

According to Björk et al. (2017), (V^π, g^π) should satisfy the following PDE system under certain regularity conditions as given in E-Companion EC.5:

$$\partial_t V^\pi + \mathcal{A}_t^\pi V^\pi + \gamma g^\pi \mathcal{A}_t^\pi g^\pi - \frac{\gamma}{2} \mathcal{A}_t^\pi (g^\pi)^2 + \lambda H(\pi_t) = 0, \quad V^\pi(T, R, X) = R, \tag{14}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

$$\partial_t g^\pi + \mathcal{A}_t^\pi g^\pi = 0, \quad g^\pi(T,R,X) = R, \tag{15}$$

where two differential operators $\mathcal{A}_t^u$ for any $u \in \mathbb{R}$ and $\mathcal{A}_t^\pi$ for any $\pi \in \mathcal{P}(\mathbb{R})$ are defined as follows:

$$\begin{aligned}\mathcal{A}_t^u \varphi(t,R,X) := &\left[r + (\mu(t,X) - r)u - \frac{1}{2}\sigma^2(t,X)u^2\right]\partial_R \varphi + m(t,X)\partial_X \varphi \\ &+ \frac{1}{2}\sigma^2(t,X)u^2 \partial_{RR}\varphi + \rho\nu(t,x)\sigma(t,x)u\partial_{RX}\varphi + \frac{1}{2}\nu^2(t,X)\partial_{XX}\varphi, \\ \mathcal{A}_t^\pi \varphi := &\int_{\mathbb{R}} \mathcal{A}_t^u \varphi \pi(du).\end{aligned} \tag{16}$$

Note that from the relaxed control perspective, the operator as given in (16) is most crucial to the system.[7]

Using the above notations, the equilibrium condition becomes

$$\pi^*(t,R,X) = \operatorname*{argmax}_{\pi \in \mathcal{P}(\mathbb{R})} \mathcal{A}_t^\pi V^* + \gamma g^* \mathcal{A}_t^\pi g^* - \frac{\gamma}{2}\mathcal{A}_t^\pi (g^*)^2 + \lambda H(\pi), \tag{17}$$

where $V^* = V^{\pi^*}$ and $g^* = g^{\pi^*}$.

Therefore, equations (14) and (15) answer our first question on policy evaluation, and the equilibrium condition (17) answers the second question on policy update. Inspired by the above observations, we propose a learning process for the equilibrium strategy as the follows:

- **Policy Evaluation Procedure:** given a policy π, compute the related value function

$$V^\pi(t,R_t,X_t) := J(t,R_t,X_t;\pi)$$

and expectation function $g^\pi(t,R_t,X_t) := \mathbb{E}_t\left[R_T^\pi | R_t, X_t\right]$;

- **Policy Update Procedure:** obtain a new policy $\tilde{\pi}$ according to optimality condition (17) with (V^π, g^π) in place of (V^*, g^*), i.e.

$$\tilde{\pi}(t,x) = \operatorname*{argmax}_{v \in \mathcal{P}(\mathbb{R})} \mathcal{A}_t^v V^\pi + \gamma g^\pi \mathcal{A}_t^v g^\pi - \frac{\gamma}{2}\mathcal{A}_t^v (g^\pi)^2 + \lambda H(v).$$

4.2. Numerical Analysis for Gaussian Mean Return Model

From now on, we shall focus on the Gaussian mean return model to discuss how to design a RL algorithm. We need to show that it really works. That is, starting from a policy π, it will converge to an equilibrium policy eventually.

Electronic copy available at: https://ssrn.com/abstract=3770818

In the RL context, it is proved that (under certain regularity condition) such learning procedure will succeed for the reward maximization problem (Sutton and Barto 2011). One key step is to establish the so-called policy improvement theorem (see, e.g. Sutton and Barto (2011) and Wang and Zhou (2020)). That is, comparing with previous policy π, one has $V^{\tilde{\pi}} \geq V^{\pi}$ for an updated policy $\tilde{\pi}$. It means that the learning process can improve the performance of the policy. Thus, it is reasonable to infer that, after sufficient iterations, the obtained policy is an optimal one, or at least, very close to that.

However, later we will see that our learning process fails to lead to a monotone iteration algorithm because what we seek is "optimality" in the equilibrium sense, and our proposed iteration lacks monotonicity in value function. Fortunately, we are able to show that under certain conditions, our iteration with a carefully selected initial guess is convergent to an equilibrium policy. Indeed, inspired by analytical form of the equilibrium policy (11), we would like to choose our initial guess for the equilibrium policy $\pi^{(0)}$ to have a constant variance and an affine function of X as its mean. To simplify our analysis, we assume that it admits the following form:

$$\pi^{(0)}(t,X) \sim \mathcal{N}\Big(\frac{X}{(1+\gamma)\sigma} - \frac{\gamma\rho\nu}{(1+\gamma)\sigma}(a_2^{(0)}(t)X + a_1^{(0)}(t)), \theta^{(0)}\Big), \tag{18}$$

where $a_2^{(0)}(t)$ and $a_1^{(0)}(t)$ are two deterministic functions of time t, and $\theta^{(0)}$ is a constant. We will show that the numerical procedure would converge as long as $a_2^{(0)}$ satisfies certain conditions.

For the initial guess as given by (18), we will show that the updated policies always keep the structure of the equilibrium policy (11). Further, we can prove that this iterative process converges locally to the desired equilibrium policy, and converges globally under the following assumption.

ASSUMPTION 1. *There exists* $M \geq 0$ *such that*

$$M \geq TC(M,T),$$

where $C(M,T) := \max\{\frac{\gamma^2}{(1+\gamma)^2}(1-\rho\nu M)^2 - 1, \frac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu T)^2 - 1\}$.

Note that if $\rho\nu$ is sufficiently small, we can verify $0 \geq TC(0,T)$ and thus Assumption 1 holds true.

We now summarize the above results as the following theorem.

Electronic copy available at: https://ssrn.com/abstract=3770818

THEOREM 2. *With the initial guess* (18)*, adopt the iteration procedure as given in Section 4.1. Denote by* $\pi^{(n)}$, $n=1,2,...$ *the sequence of updated policies. Then we have the following results.*

(i) The updated policy $\pi^{(n)}$, $n\geq 1$ *satisfies*

$$\pi^{(n)}(t,X)\sim\mathcal{N}\Big(\frac{X}{(1+\gamma)\sigma}-\frac{\gamma\rho\nu}{(1+\gamma)\sigma}(a_2^{(n)}(t)X+a_1^{(n)}(t)),\frac{\lambda}{(1+\gamma)\sigma^2}\Big),$$

where the pair $(a_1^{(n)}(t),a_2^{(n)}(t))$ *satisfies*

$$\begin{cases}\dot{a}_2^{(n)}=2\iota a_2^{(n)}+\dfrac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu a_2^{(n-1)}(t))^2-1, \quad a_2^{(n)}(T)=0,\\ \dot{a}_1^{(n)}=\iota a_1^{(n)}-\iota a_2^{(n)}+\dfrac{\gamma^2\rho\nu}{(1+\gamma)^2}(1+\rho\nu a_2^{(n-1)})a_1^{(n-1)}, \quad a_1^{(n)}(T)=0.\end{cases} \tag{19}$$

(ii) There exists $\varepsilon>0$ *such that the policy sequence* $\{\pi^{(n)}\}$ *converges to the equilibrium policy as given in (11) if* $\|a_2^{(0)}-a_2^*\|\leq\varepsilon$.

(iii) Suppose that Assumption 1 holds true, and $a_2^{(0)}(t)\in[-M,T]$ *for any* $t\in[0,T]$*. Then the policy sequence* $\{\pi^{(n)}\}$ *converges to the equilibrium policy as given in (11).*

There are two important observations in Theorem 2. First, the form in (18) is preserved in this procedure. Hence it suffices to parametrize the iterative policy through two deterministic functions $(a_2^{(n)}(t),a_1^{(n)}(t))$, rather than complicated neutral networks of functions of X, R, and t. Second, the variance of exploration equals that in the equilibrium after one step of iteration (assuming no estimation error in policy evaluation).

Part (ii) of Theorem 2 indicates a local convergence. That is, if the initial guess is close to the equilibrium policy as given in Proposition 1, then our algorithm is guaranteed to converge to the policy. For a dynamic game, a local convergence is usually the best result we could expect, as there might be multiple equilibria. However, we show that the requirement on the initial guess can be relaxed and a global convergence can be obtained. Indeed, part (iii) of Theorem 2 reveals that under some mild conditions, the iteration will converge to the desired equilibrium as long as the initial guess is bounded.

Electronic copy available at: https://ssrn.com/abstract=3770818

5. RL Algorithm Design

In this section, we will propose an RL algorithm to learn the equilibrium policy. Although the problem is in a continuous time setting in previous sections, to implement the algorithm, we have to approximate it in a discrete time setting. We then divide the time interval $[0,T]$ into N equal time intervals $[t_i,t_{i+1}],\ i=0,1,2....,N-1$, where $t_i=i\Delta t$ with $\Delta t=\frac{T}{N}$.

Let us first describe how to generate market data through simulation.

Market (simulation)

To generate the state processes by simulation, we apply the forward Euler scheme to discretize the controlled system (4) with any realized strategy u_i. That is, given market state X_i, current return R_i, and the proportion u_i invested in the risky asset at time t_i, we generate the state X_{i+1} and return R_{i+1} at time t_{i+1} in the following way:

$$\begin{aligned} X_{i+1} &= X_i+\iota(\bar{X}-X_i)\Delta t+\nu(\rho Z_{i+1}+\sqrt{1-\rho^2}\tilde{Z}_{i+1})\sqrt{\Delta t},\\ R_{i+1} &= R_i+(r+\sigma X_i u_i-\frac{1}{2}\sigma^2 u_i^2)\Delta t+\sigma\sqrt{\Delta t}Z_{i+1}, \end{aligned} \tag{20}$$

with $R_0=0$, where Z_{i+1} and $\tilde{Z}_{i+1}$ are two independent random variables drawn from standard normal distributions.

Market (real data)

When real data is available, the sequence of stock price S_i at time t_i can be directly observed. We then generate the discrete return state in terms of observed stock price:

$$R_{i+1}=R_i+u_i\frac{S_{i+1}-S_i}{S_i}+(1-u_i)r\Delta t-\frac{u_i^2}{2S_i^2}(S_{i+1}-S_i)^2. \tag{21}$$

The above construction is based on an application of Itô's formula to equation (3), namely,

$$\mathrm{d}R_t=u_t\frac{\mathrm{d}S_t}{S_t}+(1-u_t)r\mathrm{d}t-\frac{1}{2}\frac{u_t^2}{S_t^2}\mathrm{d}\langle S\rangle(t),$$

and (21) is the forward Euler discretization of the above stochastic differential equation.

Electronic copy available at: https://ssrn.com/abstract=3770818

In a real market, the market price of risk X is not directly observable. However, we may use some alternative observed data as a proxy. For example, we may follow the existing literature (e.g. Campbell and Viceira (1999) and Barberis (2000)) to approximate X by the dividend-price ratio.[8] Certainly, such approximations must lead to a bias, and we shall discuss the impact of noisy market state on our learning in Section 6.2.

Value Function and Policy parametrization

In the common practice of RL, the value function and strategy are usually parametrized with (deep) neural networks. Thanks to Proposition 1 and Theorem 2, we are able to parametrize the strategy and value functions with a simple and more explicit expression. Indeed, let $p : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ be a parametrized value function with $\mathbb{R}^K$ representing the parameter space. We then introduce the following parametrization

$$\begin{aligned} V^{\Theta}(t,R,X) &= p(\theta^{V,2}, T-t)X^2 + p(\theta^{V,1}, T-t)X + p(\theta^{V,0}, T-t) + R, \\ g^{\Theta}(t,R,X) &= p(\theta^{g,2}, T-t)X^2 + p(\theta^{g,1}, T-t)X + p(\theta^{g,0}, T-t) + R, \end{aligned} \tag{22}$$

where $\Theta = (\theta^{V,2}, \theta^{V,1}, \theta^{V,0}, \theta^{g,2}, \theta^{g,1}, \theta^{g,0}) \in \mathbb{R}^{6K}$ represents all the parameters for (V, g). One way to choose the parametrized value function p is to express it as a linear combination of several bases, e.g. $p(\theta, t) = \sum_{k=1}^{K} \theta_k \phi_k(t)$, where ϕ_k is called the feature function. The most common choices of the features are polynomials, radial basis functions, or wavelet functions. Similarly, we also parametrize the strategy as

$$\pi^{\psi,\xi} \sim \mathcal{N}\Big(W(\psi, T-t, X), \lambda\xi^2\Big),$$

where $\psi \in \mathbb{R}^{K_2}$ is the parameter in the mean, and $\lambda\xi^2$ stands for the variance of the exploration strategy. Specified parametrized functions p and W will be given later (see (25) and (26)).

Numerical analysis in Section 4.2 indicates that our iteration processes all possess the same form and converge to the equilibrium policy. As a consequence, in our implementation, we choose $W(\cdot, T-t, X)$ in the same form of the equilibrium policy in order to accelerate convergence. Compared with the widely used (deep) neural networks approximation in literature, our approximation

Electronic copy available at: https://ssrn.com/abstract=3770818

significantly saves computational costs. Our numerical results show that our approximation is already sufficiently accurate.

One may also consider to tailor the functional form in the parametrization of the value function to fit our specific model, which might simplify the resulting optimization problem. However, we do not adopt this approach due to two reasons. First, we would like to provide a general framework to show the feasibility of our algorithm that is less sensitive to a specific model. Second, the analytical form of value function is very complicated for the Gaussian mean return model, hence a specific parametrization based on the analytical form might not simplify the numerical procedure too much.

Policy Valuation Procedure

Assume that we have collected a sequence of data (t_i, R_i, X_i) with a given strategy $\pi^{\psi,\xi}$. In this policy valuation procedure, we shall choose Θ in the parameter space such that (V^Θ, g^Θ) best approximates the value functions $(V^{\pi^{\psi,\xi}}, g^{\pi^{\psi,\xi}})$. As shown in Corollary EC.1, $V^{\pi^{\psi,\xi}}$ and $g^{\pi^{\psi,\xi}}$ satisfy (14) and (15) with $\pi^{\psi,\xi}$ in place of π. From Itô formula, we see that

$$\mathbb{E}_t\left[\varphi(t+\Delta t, R^\pi_{t+\Delta t}, X_{t+\Delta t})\right] - \varphi(t, X_t, R_t) = \mathbb{E}_t\left[\int_t^{t+\Delta t} (\partial_s + \mathcal{A}^\pi_s)\varphi(s, R^\pi_s, X_s) ds\right].$$

Since $g^{\pi^{\psi,\xi}}$ solves (15), it implies that $g^{\pi^{\psi,\xi}}(t, X_t, R_t)$ is a martingale. If g^Θ is close to $g^{\pi^{\psi,\xi}}$, it is natural to think that g^Θ is close to be a martingale.

Define

$$C_i^2(\Theta, \psi, \xi) := \frac{g^\Theta(t_{i+1}, R_{i+1}, X_{i+1}) - g^\Theta(t_i, R_i, X_i)}{\Delta t} \tag{23}$$

and

$$\begin{aligned} C_i^1(\Theta, \psi, \xi) :=& \frac{V^\Theta(t_{i+1}, R_{i+1}, X_{i+1}) - V^\Theta(t_i, R_i, X_i)}{\Delta t} \\ &+ \gamma g^\Theta(t_i, R_i, X_i)\frac{g^\Theta(t_{i+1}, R_{i+1}, X_{i+1}) - g^\Theta(t_i, R_i, X_i)}{\Delta t} \\ &- \frac{\gamma}{2}\frac{(g^\Theta)^2(t_{i+1}, R_{i+1}, X_{i+1}) - (g^\Theta)^2(t_i, R_i, X_i)}{\Delta t} + \lambda H(\pi^{\psi,\xi}_{t_i}). \end{aligned} \tag{24}$$

The value function can be approximated by choosing Θ that minimizes

$$\sum_{i=0}^{N-1} \left(C_i^1(\Theta, \psi, \xi)\right)^2 + \left(C_i^2(\Theta, \psi, \xi)\right)^2.$$

Electronic copy available at: https://ssrn.com/abstract=3770818

C_i^1 and C_i^2 are also called the temporal difference error (TD error) or the Bellman error that has been used in Doya (2000) and Wang and Zhou (2020) for dynamic optimization problems. In the algorithm, instead of directly finding the exact minimizer, we will use the stochastic gradient based algorithm to solve this minimization problem for one step and then start to iterate. The squared error is widely used in many RL algorithms (cf. Doya (2000) and Sutton and Barto (2011)) due to the convention that the square error minimizer might be a reasonable approximation of the condition expectation.[9]

Policy Update Procedure

In the policy update procedure, we shall update policy according to the optimality condition (17) for the given function (V^Θ, g^Θ). Define

$$L(\psi,\xi;t,R,X) := \mathcal{A}^{\pi^{\psi,\xi}}V^\Theta + \gamma g^\Theta \mathcal{A}^{\pi^{\psi,\xi}} g^\Theta - \frac{\gamma}{2}\mathcal{A}^{\pi^{\psi,\xi}}(g^\Theta)^2 + \lambda H(\pi^{\psi,\xi}).$$

Ideally, the updated policy should maximize $L(\psi,\xi;t,R,X)$ over all possible values of (t,R,X). Since we only have the data (t_i,R_i,X_i), we shall try to maximize $\sum_{i=0}^{N-1} L(\psi,\xi;t_i,R_i,X_i)$ instead. Once again, we plan to use the gradient descent method to find the optimal solution. However, we cannot compute the derivative directly for the current parameter (ψ,ξ) since the operator $\mathcal{A}$ is unknown. To approximate the derivative, we use the smooth functional method that has been used by Prashanth and Ghavamzadeh (2013) for risk-sensitive problems (see Bhatnagar et al. (2013) for an introduction). The method is motivated by the following observation. Let δ_κ be the Gaussian intensity with variance κ^2. For any continuous differentiable function L, we see that

$$\int_{\mathbb{R}} \delta_\kappa(x-z)\nabla L(z)dz = \int_{\mathbb{R}} \nabla\delta_\kappa(x-z)L(z) = \frac{1}{\kappa}\int_{\mathbb{R}} z'\delta_1(z')L(x-\gamma z')dz'.$$

When $\kappa \to 0$, the left hand side of the above equation converges to $\nabla L(x)$. Thus, a reasonable estimation of $\nabla L(x)$ would be

$$\nabla L(x) \approx \frac{\Delta}{\kappa}(L(x+\kappa\Delta)).$$

Electronic copy available at: https://ssrn.com/abstract=3770818

One can further reduce the variance by using the following approximation

$$\nabla L(x) \approx \frac{\Delta}{\kappa}(L(x+\kappa\Delta) - L(x)).$$

To apply the smooth functional method, we need to compute $L(\tilde{\psi}, \tilde{\xi}; t_i, R_i, X_i)$ for a set of perturbed parameters $(\tilde{\psi}, \tilde{\xi})$ as well as $L(\psi, \xi; t_i, R_i, X_i)$ for the current policy parameters (ψ, ξ). From discussion in the previous part, it is natural to approximate $L(\psi, \xi; t_i, R_i, X_i)$ as

$$L(\psi, \xi; t_i, R_i, X_i) \approx C_i(\Theta, \psi, \xi).$$

For the perturbed strategy, we shall use the TD error once again to approximate $L(\tilde{\psi}, \tilde{\xi}; t_i, R_i, X_i)$. For that purpose, at time t_i, besides using current strategy $\pi^{\psi,\xi}$, we shall also use another perturbed strategy $\pi^{\tilde{\psi},\tilde{\xi}}$ to obtain another return $\tilde{R}_{i+1}$. Then, $L(\tilde{\psi}, \tilde{\xi}; t_i, R_i, X_i)$ is approximated as

$$\begin{aligned} L(\tilde{\psi}, \tilde{\xi}; t_i, R_i, X_i) \approx & \frac{V^\Theta(t_{i+1}, \tilde{R}_{i+1}, X_{i+1}) - V^\Theta(t_i, R_i, X_i)}{\Delta t} \\ & + \gamma g^\Theta(t_i, R_i, X_i) \frac{g^\Theta(t_{i+1}, \tilde{R}_{i+1}, X_{i+1}) - g^\Theta(t_i, R_i, X_i)}{\Delta t} \\ & - \frac{\gamma}{2} \frac{(g^\Theta)^2(t_{i+1}, \tilde{R}_{i+1}, X_{i+1}) - (g^\Theta)^2(t_i, R_i, X_i)}{\Delta t} + \lambda H(\pi_{t_i}^{\tilde{\psi},\tilde{\xi}}). \end{aligned}$$

Having this, we can approximate the derivative and apply the gradient descent method. Finally, in our implementation, we apply the Adam algorithm introduced by Kingma and Ba (2014) in policy update procedure to approximate the equilibrium strategy.

It should be emphasized that the policy update procedure in our algorithm is very different from the algorithm for learning the pre-committed strategy in Wang and Zhou (2020), where their algorithm can be guaranteed to improve the strategy monotonically. However, we do not have such monotonic pattern for an equilibrium solution, since there is no optimality in terms of value functions. On the other hand, we compute the gradient with respect to the parameters in the policy (known as the policy gradient) to update our policy, according to optimality in the equilibrium condition (17).

Combining all these together, we propose the following algorithm.

Electronic copy available at: https://ssrn.com/abstract=3770818

Algorithm 1 Exploring Equilibrium Strategy

Input: Market data, riskless interest rate r, learning rate α, investment horizon T, discretization Δt, exploration rate λ, number of iterations M, smoothing functional parameter κ, risk aversion γ, the parametrized value function $p(\theta,t)$, and $W(\psi,t,X)$

Initialize θ,ψ,ξ

for $k=1$ to M **do**

for $i=0$ to $N-1$ **do**

Sample $\Delta^{1,i}\sim\mathcal{N}(0,I_{2K}),\Delta^{2,i}\sim\mathcal{N}(0,1)$, $\varepsilon\sim\mathcal{N}(0,1)$

Compute current investment decision $u=W(\psi,T-t_i,X_i)+\xi\sqrt{\lambda}\varepsilon$

Compute perturbed investment decision $\tilde{u}=W(\psi+\kappa\Delta^{1,i},T-t_i,X_i)+(\xi+\kappa\Delta^{2,i})\sqrt{\lambda}\varepsilon$

Sample R_{i+1},X_{i+1} from *Market* with decision u and current state (R_i,X_i)

Sample $\tilde{R}_{i+1}$ from *Market* with decision $\tilde{u}$ and current state (R_i,X_i) at the same time

end for

Policy Valuation Procedure

Compute the TD error based on (24) and (23).

Compute $\Delta\theta^{V,j}=\sum_i C_i^1(\Theta,\psi,\xi)\partial_{\theta^{V,j}}(V^\Theta(t_{i+1},R_{i+1},X_{i+1})-V^\Theta(t_i,R_i,X_i)),j=0,1,2$

Compute $\Delta\theta^{g,j}=\sum_i C_i^2(\Theta,\psi,\xi)\partial_{\theta^{g,j}}(g^\Theta(t_{i+1},R_{i+1},X_{i+1})-g^\Theta(t_i,R_i,X_i)),j=0,1,2$

Update $\theta^{V,j}\leftarrow\theta^{V,j}-\alpha\Delta\theta^{V,j}$

Update $\theta^{g,j}\leftarrow\theta^{g,j}-\alpha\Delta\theta^{g,j}$

Policy Update Procedure

Compute the TD error $C_i^1(\Theta,\psi,\xi)$ with new Θ and

$$\begin{aligned}\tilde{C}_i^1(\Theta,\tilde{\psi},\tilde{\xi})=&\frac{V^\Theta(t_{i+1},\tilde{R}_{i+1},X_{i+1})-V^\Theta(t_i,R_i,X_i)}{\Delta t}\\&+\gamma g^\Theta(t_i,R_i,X_i)\frac{g^\Theta(t_{i+1},\tilde{R}_{i+1},X_{i+1})-g^\Theta(t_i,R_i,X_i)}{\Delta t}\\&-\frac{\gamma}{2}\frac{(g^\Theta)^2(t_{i+1},\tilde{R}_{i+1},X_{i+1})-(g^\Theta)^2(t_i,R_i,X_i)}{\Delta t}+\lambda H(\pi_{t_i}^{\tilde{\psi},\tilde{\xi}}).\end{aligned}$$

$\Delta\psi=\sum_{i=0}^{N-1}\frac{\Delta^{1,i}}{\kappa}(\tilde{C}_i^1(\Theta,\tilde{\psi},\tilde{\xi})-C_i^1(\Theta,\psi,\xi))$

$\Delta\xi=\sum_{i=0}^{N-1}\frac{\Delta^{2,i}}{\kappa}(\tilde{C}_i^1(\Theta,\tilde{\psi},\tilde{\xi})-C_i^1(\Theta,\psi,\xi))$

Electronic copy available at: https://ssrn.com/abstract=3770818

Update ψ, ξ with $\Delta\psi$ and $\Delta\xi$ by Adam algorithm

end for

6. Numerical Results

Now we conduct numerical experiments with simulated data to demonstrate our algorithm. As a special case of the Gaussian mean return model, the numerical results under Black-Scholes model are illustrated in E-Companion EC.2.

6.1. Gaussian Mean Return Model

We use the parameters estimated in Wachter (2002): $\rho = -0.93,\ r = 0.017,\ \sigma = 0.15,\ X_0 = \bar{X} = 0.273,\ \iota = 0.27$, and $\nu = 0.065$ to generate data. For the algorithm, we set $T = 1,\ \Delta t = \frac{1}{250},\ \alpha = 0.001,\ M = 50000$ and $\gamma = 2$. For the hyper-parameters in the Adam method, we use the values recommended by Kingma and Ba (2014). The parametrized value function $p(\theta, t)$ is chosen to be

$$p(\theta, t) = \theta_0 + \theta_1 t + \theta_2 t^2. \tag{25}$$

The parametrized function W of the strategy is chosen to be

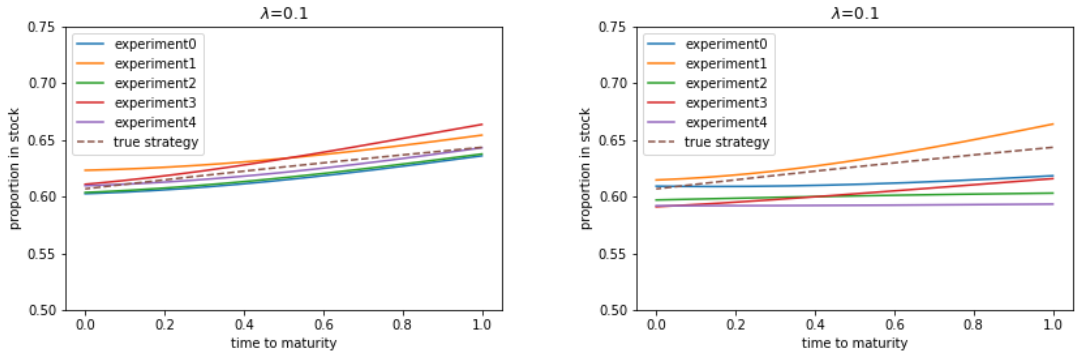
$$\begin{aligned} W(\psi, T-t, X) =& \psi_1 \frac{e^{2\psi_2(T-t)} - 1}{(\psi_3 + \psi_2)(e^{\psi_2(T-t)} - 1) + 2q} x + \psi_4 x \\ &+ \psi_5 \frac{(e^{\psi_2(T-t)} - 1)^2}{\psi_2[(\psi_3 + \psi_2)(e^{2\psi_2(T-t)} - 1) - 2\psi_2]} \end{aligned} \tag{26}$$

with $\psi = (\psi_1, \psi_2, \psi_3, \psi_4, \psi_5)$. Note that, for $\psi^* = (\psi_1^*, \psi_2^*, \psi_3^*, \psi_4^*, \psi_5^*) = (C_1, q, b, \frac{1}{(1+\gamma)\sigma}, C_2)$ with the coefficients given in Proposition 1, $W(\psi^*, T-t, X)$ is the theoretical true equilibrium strategy. Moreover, due to the terminal condition in (14) and (15), we further require that $\theta_0^{V,i} = 0,\ \theta_0^{g,i} = 0$ for $i = 0, 1, 2$. The initial values for parameters of value functions are all set to be 0. As we have proved in Section 4, the convergence is guaranteed when the initial value of ψ is properly chosen. The initial value are chosen as $\psi_1 = \psi_3 = \psi_5 = 0$ and $\psi_2 = \frac{1}{1+\gamma}$. The parameter ψ_4 is important, which is associated with the corresponding myopic strategy. Observe that its true value $\psi_4^* = \frac{1}{(1+\gamma)\sigma}$, while the optimal value of ξ is $\xi^* = \frac{1}{(1+\gamma)\sigma^2}$. Thus, the relation between these two optimal values is $\psi_4^* = \frac{\sqrt{\xi^*}}{\sqrt{1+\gamma}}$. This motivates us to introduce a pre-training stage to get a good initial value of

Electronic copy available at: https://ssrn.com/abstract=3770818

ψ_4. In the pre-training stage, we initialize $\psi_4 = \frac{4}{1+\gamma}$ and iterate for 25000 steps. Then, we obtain the trained optimal parameter $\tilde{\xi}$. After the pre-training stage, we start the training with initial value for ψ_4 as $\frac{\sqrt{\tilde{\xi}}}{\sqrt{1+\gamma}}$. We use $M = 50000$ sample paths to train the model and pick up 25000 paths randomly for the pre-train stage. To eliminate the noises introduced by the algorithm, we train the model for 5 times with the same but shuffled dataset starting with the same initial value. At last, we use the average parameters as our final result. Furthermore, to see whether the algorithm is stable or not, we conduct the experiment for five times. For each time, $M = 50000$ sample paths are generated for the training.

In Figure 1(a), we plot the mean value of the proportion invested in the risky asset for the strategy when market state $X = \bar{X} = 0.273$ at the current moment. Overall, the output of our algorithm is very close to the theoretical value.



(a) Perfect observation of X (b) Noisy observation of X

Figure 1 **The strategy calculated by our algorithm at different time when $X = \bar{X}$ versus the theoretical solution.** Panel (a) is computed given perfect observation of market state X, while Panel (b) is computed by assuming we have a noisy observation of X. In this example, the data is simulated with parameters $\rho = -0.93$, $r = 0.017$, $\sigma = 0.15$, $X_0 = \bar{X} = 0.273$, $\iota = 0.27$, and $\nu = 0.065$, $T = 1$, $\Delta t = \frac{1}{250}$, $\alpha = 0.001$, $M = 50000$, $\gamma = 2$, and we fix $\lambda = 0.1$. In panel (b), the noise in the observation of X has standard deviation 0.05.

To further compare the learned strategy with the true strategy, we consider the relative error between them for different (t, x). We choose $t_i = \frac{iT}{N}$ and $X_j = \bar{X} + j * 0.01$ with $i = 0, 1, 2, ..., N = 250$

Electronic copy available at: https://ssrn.com/abstract=3770818

and $j = 0, \pm 1, \pm 2, ..., \pm 10$. Then, for a parameter ψ, the average relative error is defined as

$$\text{Relative error} = \frac{1}{21(N+1)} \sum_{i,j} \frac{|W(\psi, T-t_i, X_j) - W(\psi^*, T-t_i, X_j)|}{|W(\psi^*, T-t_i, X_j)|}.$$

Table 1 shows the average relative error for the five experiments. It reflects the average error across all time steps and all possible state variables. If we average the strategy learned from five experiments, we shall reduce the test error since the noise in the algorithm is canceled out.

6.2. Noisy Market State

To test whether our algorithm is stable with noise, we consider the case that the true value of market state X cannot be measured accurately and what the investor observes is a noise data, i.e. the true value plus a random noise. To simulate this scenario, we adjust the market simulator in the following way. At time t_i, the market state and return still evolve as in (20). But, the simulator will not generate the true value X_i directly. Instead, it returns the noise value $\tilde{X}_i$ defined as

$$\tilde{X}_i = X_i + 0.05 * \varepsilon_i,$$

where ε_i is an i.i.d. random variable drawn from a standard normal distribution. Note that the value of X_i will be close to $\bar{X} = 0.273$. Thus, the noise added is not negligible. This setting mimics the case where investors may infer the market price of risk from other data with estimation error.

We apply our algorithm for this case with the same hyper parameters and initialization as before. We also train our model in five experiments for different values of λ. To compare the learned strategy with the theoretical one, we plot the proportion for investment when $X = \bar{X}$ in Figure 1(b). Clearly, the deviation of our strategy from the theoretical one increases compared to the case with perfect observation. Moreover, we notice that error would increase more significantly as the investment horizon grows, which may be attributed to the fact that the observation noise accumulates with time.

The relative error over all periods and all states of the learned strategy is shown in Table 1. It almost doubles the error under perfect observation. More importantly, a noisy observation in

Electronic copy available at: https://ssrn.com/abstract=3770818

market state creates a new source of error that cannot be easily eliminated by simple average. Despite that, the overall error is around 3% and remains acceptable.

Table 1 Relative error for our strategies against the theoretical value. In this example, the data is simulated with parameters $\rho = -0.93,\ r = 0.017,\ \sigma = 0.15,\ X_0 = \bar{X} = 0.273,\ \iota = 0.27$, and $\nu = 0.065$, $T = 1$, $\Delta t = \frac{1}{250},\ \alpha = 0.001,\ M = 50000$, $\gamma = 2$, and we fix $\lambda = 0.1$. In left column, the noise in the observation of X has standard deviation 0.05.

	Perfect observation of X	Noisy observation of X
Experiment 0	1.47%	2.21%
Experiment 1	1.59%	1.46%
Experiment 2	1.20%	4.04%
Experiment 3	1.47%	3.66%
Experiment 4	0.53%	5.31%
Average Strategy	0.35%	2.73%

7. Conclusions

Inspired by Dai et al. (2020) and Wang et al. (2019), we incorporate an entropy-regularizer into a dynamic mean-variance criterion for portfolio's log return, in order to learn an equilibrium mean-variance strategy in terms of the reinforcement learning technique. Under an incomplete market setting, we obtain a semi-analytical, exploratory, equilibrium mean-variance strategy that turns out to follow a Gaussian distribution.

For algorithm design and implementation, we focus on a Gaussian mean return model for which a closed-form, exploratory, equilibrium mean-variance strategy is available. To find the equilibrium strategy using reinforcement learning, we design a policy iteration procedure with a thoroughly selected initial guess such that the iteration policies possess the same structure as the targeted equilibrium strategy. It should be highlighted that dynamic programming principle fails with an equilibrium solution and the corresponding policy updates may not improve the original policy. However, we are still able to prove the convergence of our algorithm under mild conditions, thanks

Electronic copy available at: https://ssrn.com/abstract=3770818

to a thorough design of the policy iteration procedure. Numerical experiments are given to demonstrate our algorithm.

There are some open questions along this direction, such as the algorithm design for a general incomplete markets, how to incorporate transaction costs into this framework, an empirical study of our algorithm, etc. We will leave them for future study.

Endnotes

1. Recently, He and Jiang (2019) study an equilibrium strategy for a constrained mean-variance formulation.

2. Time-inconsistency is a broad issue widely discussed in economics and decision science; see Strotz (1955).

3. From this angle, the unique Brownian motion discussed in the exploratory formulation of Wang et al. (2019) is different from the original Brownian motion.

4. It is also known as the temperature constant that measures the trade-off between exploitation and exploration.

5. To further illustrate the difference of the variance term between the pre-committed strategy and the equilibrium strategy under the exploration framework, we consider the same objective function and control variables in E-Companion EC.3 and we also find that the equilibrium policy has a constant variance.

6. In the stochastic control literature, only the expectation of the payoff function associated with the optimal policy is called the "value function". However, we follow the convention in RL literature to call the expectation associated with any policy the "value function", and use a superscript to indicate the policy.

7. This operator is known as the "infinitesimal generator". An infinitesimal generator associated with a controlled process R_t^π under the policy π and initial state $R_0^\pi = R$ is an operator $\mathcal{A}^\pi$ such that $\varphi(R_t^\pi) - \varphi(R) - \int_0^t \int_{\mathbb{R}} \mathcal{A}_s^u \varphi(R_s^\pi) \pi_s(du) ds$ is a martingale for any smooth φ. We often omit the subscript t and denote this operator by $\mathcal{A}^u$ or $\mathcal{A}^\pi$ when it does not cause any confusion.

Electronic copy available at: https://ssrn.com/abstract=3770818

8. In fact, there is a large body of literature in finance investigating the predictability of stock returns, e.g. Ang and Bekaert (2007) and Cochrane (2008). They provide comprehensive examinations of several factors that can be regraded as a proxy for our state variable X. In our framework, since we allow X to be a stochastic process, it characterizes the noise and limited predictability of the observed factors. Wachter (2002) also justifies the Gaussian mean return model by claiming that the dividend-price ratio used in Campbell and Viceira (1999) and Barberis (2000) is a commonly used factor.

9. Even though one can show that the minimizer of the above accumulative square loss is not consistent with our desired function, it is still adopted in some RL literature in the continuous-time setup (e.g. Doya (2000) and Wang and Zhou (2020)).

Acknowledgments

Dai acknowledges the supports of Singapore MOE AcRF grants R-146-000-306-114 and R-703-000-032-112.

References

Aït-Sahali Y, Brandt MW (2001) Variable selection for portfolio choice. The Journal of Finance 56(4):1297–1351.

Ang A, Bekaert G (2007) Stock return predictability: Is it there? The Review of Financial Studies 20(3):651–707.

Azar MG, Gómez V, Kappen B (2011) Dynamic policy programming with function approximation. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 119–127.

Barberis N (2000) Investing for the long run when returns are predictable. The Journal of Finance 55(1):225–264.

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

Bhatnagar S, Prasad H, Prashanth L (2013) Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods (Springer).

Electronic copy available at: https://ssrn.com/abstract=3770818

Björk T, Khapko M, Murgoci A (2017) On time-inconsistent stochastic control in continuous time. Finance and Stochastics 21(2):331–360.

Björk T, Murgoci A, Zhou XY (2014) Mean–variance portfolio optimization with state-dependent risk aversion. Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics 24(1):1–24.

Black F, Litterman R (1991) Asset allocation combining investor views with market equilibrium. Journal of Fixed Income 1(2):7–18.

Brandt MW (1999) Estimating portfolio and consumption choice: A conditional Euler equations approach. The Journal of Finance 54(5):1609–1645.

Campbell JY, Viceira LM (1999) Consumption and portfolio decisions when expected returns are time varying. The Quarterly Journal of Economics 114(2):433–495.

Campbell JY, Viceira LM (2002) Strategic Asset Allocation: Portfolio Choice for Long-term Investors (Clarendon Lectures in Economic).

Cochrane JH (2008) The dog that did not bark: A defense of return predictability. The Review of Financial Studies 21(4):1533–1575.

Cover TM, Thomas JA (2012) Elements of Information Theory (John Wiley & Sons).

Dai M, Jin H, Kou S, Xu Y (2020) A dynamic mean-variance analysis for log returns. Management Science .

Deisenroth MP, Neumann G, Peters J, et al. (2013) A survey on policy search for robotics. Foundations and Trends® in Robotics 2(1–2):1–142.

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

Fleming WH, Nisio M (1984) On stochastic relaxed control for partially observed diffusions. Nagoya Mathematical Journal 93:71–108.

Fox R, Pakman A, Tishby N (2015) Taming the noise in reinforcement learning via soft updates. arXiv preprint arXiv:1512.08562 .

Gao X, Chan L (2000) An algorithm for trading and portfolio management using q-learning and sharpe ratio maximization. Proceedings of the International Conference on Neural Information Processing, 832–837.

Electronic copy available at: https://ssrn.com/abstract=3770818

Guéant O, Manziuk I (2019) Deep reinforcement learning for market making in corporate bonds: beating the curse of dimensionality. Applied Mathematical Finance 26(5):387–452.

He X, Jiang Z (2019) Mean-variance portfolio selection with dynamic targets for expected terminal wealth. Chinese University of Hong Kong .

Hendricks D, Wilcox D (2014) A reinforcement learning extension to the almgren-chriss framework for optimal trade execution. 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 457–464 (IEEE).

Jin O, El-Saawy H (2016) Portfolio management using reinforcement learning. Technical Report, Stanford University .

Kim TS, Omberg E (1996) Dynamic nonmyopic portfolio behavior. The Review of Financial Studies 9(1):141–161.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research 17(1):1334–1373.

Li D, Ng WL (2000) Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. Mathematical Finance 10(3):387–406.

Liu J (2007) Portfolio selection in stochastic environments. The Review of Financial Studies 20(1):1–39.

Luenberger DG (1998) Investment Science (Oxford University Press: New York).

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

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

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529.

Moody J, Saffell M (2001) Learning to trade via direct reinforcement. IEEE transactions on neural Networks 12(4):875–889.

Electronic copy available at: https://ssrn.com/abstract=3770818

Moody J, Wu L, Liao Y, Saffell M (1998) Performance functions and reinforcement learning for trading systems and portfolios. Journal of Forecasting 17(5-6):441–470.

Nachum O, Norouzi M, Xu K, Schuurmans D (2017) Bridging the gap between value and policy based reinforcement learning. Advances in Neural Information Processing Systems, 2775–2785.

Neuneier R (1996) Optimal asset allocation using adaptive dynamic programming. Advances in Neural Information Processing Systems, 952–958.

Nevmyvaka Y, Feng Y, Kearns M (2006) Reinforcement learning for optimized trade execution. Proceedings of the 23rd International Conference on Machine Learning, 673–680 (ACM).

Prashanth L, Ghavamzadeh M (2013) Actor-critic algorithms for risk-sensitive mdps. Advances in Neural Information processing systems, 252–260.

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

Ritter G (2017) Machine learning for trading. Working Paper. Available at SSRN 3015609 .

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

Sato M, Kobayashi S (2000) Variance-penalized reinforcement learning for risk-averse asset allocation. International Conference on Intelligent Data Engineering and Automated Learning, 244–249 (Springer).

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484.

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. (2017) Mastering the game of go without human knowledge. Nature 550(7676):354.

Sobel MJ (1982) The variance of discounted markov decision processes. Journal of Applied Probability 19(4):794–802.

Strotz RH (1955) Myopia and inconsistency in dynamic utility maximization. The Review of Economic Studies 23(3):165–180.

Electronic copy available at: https://ssrn.com/abstract=3770818

Sutton RS, Barto AG (2011) Reinforcement learning: An Introduction (Cambridge, MA: MIT Press).

Tamar A, Mannor S (2013) Variance adjusted actor critic algorithms. arXiv preprint arXiv:1310.3697 .

Todorov E (2007) Linearly-solvable markov decision problems. Advances in Neural Information processing systems, 1369–1376.

Tokic M (2010) Adaptive ε-greedy exploration in reinforcement learning based on value differences. Annual Conference on Artificial Intelligence, 203–210 (Springer).

Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. Thirtieth AAAI Conference on Artificial Intelligence.

Wachter JA (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, Zariphopoulou T, Zhou XY (2019) Reinforcement learning in continuous time and space: A stochastic control approach. Journal of Machine Learning Research .

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

Watkins CJ, Dayan P (1992) Q-learning. Machine Learning 8(3-4):279–292.

Zhou XY (1992) On the existence of optimal relaxed controls of stochastic partial differential equations. SIAM Journal on Control and Optimization 30(2):247–261.

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

Ziebart BD, Maas AL, Bagnell JA, Dey AK (2008) Maximum entropy inverse reinforcement learning. AAAI, volume 8, 1433–1438 (Chicago, IL, USA).

Electronic copy available at: https://ssrn.com/abstract=3770818

E-Companion of Learning Equilibrium Mean-Variance Strategy

EC.1. Design for Time-decaying Exploration

In order to induce a time-decaying variance in our framework, one natural choice is to provide a time-decaying reward for exploration. For example, the objective functional (5) can be modified as:

$$J(t,R_t,X_t;\pi):=\mathbb{E}_t\left[R_T^\pi+\lambda\int_t^T e^{-\delta s}H(\pi_s)ds\right]-\frac{\gamma}{2}\mathrm{Var}_t\left[R_T^\pi\right], \tag{EC.1}$$

where a discount factor $e^{-\delta(s-t)}$ is introduced so that the future exploration will be rewarded less. $\delta\geq 0$ captures the rate of decay in the variance term. When $\delta=0$, it reduces to (5) .

Under the Black-Scholes model, the equilibrium strategy now becomes

$$\pi_t^*\sim\mathcal{N}\Big(\frac{\mu-r}{(1+\gamma)\sigma^2},\frac{\lambda}{(1+\gamma)\sigma^2}e^{-\delta t}\Big). \tag{EC.2}$$

And for the general model we discussed, we have that

$$\pi^*(t,X)\sim\mathcal{N}\Big(\frac{\mu(t,X)-r-\rho\gamma\sigma(t,X)\nu(t,X)\partial_X h^*(t,X)}{(1+\gamma)\sigma^2(t,X)},\frac{\lambda}{(1+\gamma)\sigma^2(t,X)}e^{-\delta t}\Big). \tag{EC.3}$$

The proof about the above conclusions will be parallel to the case $\delta=0$ in EC.2 and EC.5, and hence will be omitted.

EC.2. Black-Scholes Model

In this section, we consider the complete market case with constant market parameters, i.e. $\mu_t\equiv\nu,\sigma_t\equiv\sigma$. The entropy regularized portfolio choice problem with objective functional (5) yields a closed form solution in the complete market as follows.

Proposition EC.1. *Under the complete market setting with constant market parameters, an equilibrium strategy is given by*

$$\pi_t^*\sim\mathcal{N}\Big(\frac{\mu-r}{(1+\gamma)\sigma^2},\frac{\lambda}{(1+\gamma)\sigma^2}\Big). \tag{EC.4}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

Proof of Proposition EC.1 For any deterministic strategy π, we can get that

$$\mathbb{E}_t[R_T] = R_t + \int_t^T r + (\mu - r)\int_{\mathbb{R}} u\pi_s^*(du) - \frac{1}{2}\int_{\mathbb{R}} u^2\pi_s^*(du)ds$$

and $\mathrm{Var}_t[R_T] = \int_t^T \sigma^2 \int_{\mathbb{R}} u^2\pi_s(du)ds$. Since both π^* defined in (EC.4) and its perturbation $\pi^{h,v}$ are deterministic, we shall have that

$$\begin{aligned}J(t,X_t,R_t;\pi^{h,v}) =& J(t,X_t,R_t;\pi^*) - \int_t^{t+h}\left(r + (\mu-r)\int_{\mathbb{R}} u\pi_s^*(du) - \frac{1+\gamma}{2}\int_{\mathbb{R}} u^2\pi_s^*(du) + \lambda H(\pi_s^*)\right)ds \\ &+ \int_t^{t+h}\left(r + (\mu-r)\int_{\mathbb{R}} uv(du) - \frac{1+\gamma}{2}\int_{\mathbb{R}} u^2 v(du) + \lambda H(v)\right)ds.\end{aligned}$$

Then, it holds that

$$\begin{aligned}&\liminf_{h\to 0^+}\frac{J(t,R_t,X_t;\pi^*) - J(t,R_t,X_t;\pi^{h,v})}{h} \\ =&(\mu-r)\int_{\mathbb{R}} u\pi_t^*(du) - \frac{1+\gamma}{2}\int_{\mathbb{R}} u^2\pi_t^*(du) + \lambda H(\pi_t^*) - \left((\mu-r)\int_{\mathbb{R}} uv(du) - \frac{1+\gamma}{2}\int_{\mathbb{R}} u^2 v(du) + \lambda H(v)\right)\end{aligned}$$

The conclusion of the theorem can be deduced from two facts. One is that of all probability distribution over the reals with a specific mean α and variance β, the normal distribution is the one with the maximal entropy which equals to $\frac{1}{2}\log(2\pi e\beta)$. The other is that $(\hat{\alpha},\hat{\beta}) = (\frac{\mu-r}{(1+\gamma)\sigma^2}, \frac{\lambda}{(1+\gamma)\sigma^2})$ attains the maxima of the function $(\mu-r)\alpha - \frac{1+\gamma}{2}\sigma^2(\alpha^2+\beta) + \frac{\lambda}{2}\log(2\pi e\beta)$. □

A remarkable feature of the derived equilibrium strategy is that its mean coincides with that of the original, non-exploratory problem (see Dai et al. (2020)), whereas the temperature parameter λ and the volatility σ determine the variance. For fixed λ, the smaller the volatility the bigger the variance of the equilibrium strategy. It seems not natural at first glance, since intuitively, less exploration is needed when the market noise is small. But, from the $d\bar{B}_t$ term in (4), we see that, for the equilibrium strategy, the overall variance introduced by the exploration into the system is constant independent of the value of volatility. In other words, the exploration is constant regardless of the market.

We now test our algorithm with simulated data. In the numerical experiment, we set $\mu = 0.1$, $\sigma = 0.3$, $r = 0.017$, and $\gamma = 2$. For the algorithm, we set $T = 1$, $\Delta t = \frac{1}{250}$, $\alpha = 0.001$, $M = 50000$. And λ is taken as 0, 0.1, 1 to examine the impact of exploration level. As shown in Appendix EC.2, the equilibrium policy is

$$\pi_t^* \sim \mathcal{N}\Big(\frac{\mu-r}{(1+\gamma)\sigma^2}, \frac{\lambda}{(1+\gamma)\sigma^2}\Big).$$

Electronic copy available at: https://ssrn.com/abstract=3770818

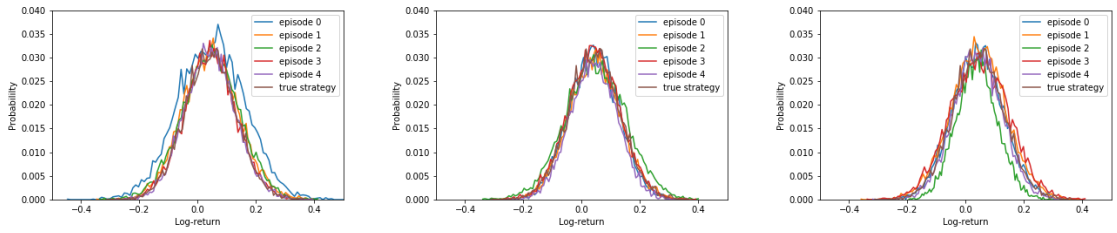
From the previous analysis, we parametrize the strategy and value function as

$$\pi^{\psi,\xi} \sim N(\psi, \xi^2), \quad V(t,R) = R + \theta^V(T-t), \quad g(t,R) = R + \theta^g(T-t).$$

We use 20000 paths as training samples and another 10000 path to get the empirical distribution of the terminal log-return for the obtained strategy as testing samples. We repeat the learning procedure 5 times and also compare the learned strategy with the true strategy.

The result is presented in Figure EC.1. Roughly speaking, the log-returns under our strategy are distributed similar to that under the theoretical equilibrium strategy. Left and right tails are symmetric and have similar rates of decay as a Gaussian distribution. It implies that our method could approximate theoretical solution accurately.

To further illustrate the result, we present the mean and standard deviation of the proportion ψ for different values of λ in Table EC.1. We can see that the mean of the proportion of wealth invested in stocks is very close to the theoretical value and the standard deviation of the mean value (not the standard deviation in the exploration) is very small, which demonstrates the convergence of our algorithm.



(a) $\lambda = 0$ (b) $\lambda = 0.1$ (c) $\lambda = 1$

Figure EC.1 **The histogram of the log-return of the portfolio under the strategy according to our algorithm** when the stock price is generated based on the Black-Scholes model with different values of exploration parameter λ. Other parameters are taken as $\mu = 0.1$, $\sigma = 0.3$, $r = 0.017$, and $\gamma = 2$, $T = 1$, $\Delta t = \frac{1}{250}$, $\alpha = 0.001$, $M = 50000$.

Electronic copy available at: https://ssrn.com/abstract=3770818

Table EC.1 Sample average and standard deviation of the mean value of the policy generated by our algorithm. Other parameters are taken as $\mu = 0.1$, $\sigma = 0.3$, $r = 0.017$, and $\gamma = 2$, $T = 1$, $\Delta t = \frac{1}{250}$, $\alpha = 0.001$, $M = 50000$.

Exploration Parameter	Mean of ψ	Standard Deviation of ψ
$\lambda = 0$	0.321	0.030
$\lambda = 0.1$	0.296	0.025
$\lambda = 1$	0.292	0.033
Theoretical Value of $\psi = \frac{\mu - r}{(1+\gamma)\sigma^2} = 0.296$		

EC.3. Comparison between the Pre-committed Policy and the Equilibrium Policy

To shed some light on the difference between the pre-committed policy and the equilibrium in the exploratory framework, we adopt the same setting and objective functional as in Wang and Zhou (2020) to make comparison, especially the difference in the exploratory aspect (variance).

We use the same control variables and state process the discounted wealth process $\bar{W}_t = e^{-rt}W_t$ as Wang and Zhou (2020). The control policy is denoted by π that stands for the distribution of the discounted wealth invested in stocks. More specifically, if the realization of the control policy is denoted by u_t, then

$$\mathrm{d}\bar{W}_t^u = (\mu - r)u_t\mathrm{d}t + \sigma u_t \mathrm{d}B_t.$$

And the equivalent exploratory dynamics is

$$\mathrm{d}\bar{W}_t^\pi = (\mu - r)\int_{\mathbb{R}} u\pi_t(u)\mathrm{d}u\mathrm{d}t + \sigma\sqrt{\int_{\mathbb{R}} u^2\pi_t(u)\mathrm{d}u}\mathrm{d}B_t.$$

The objective functional is

$$\mathbb{E}_t[\bar{W}_T^\pi] - \frac{\gamma}{2}Var_t[\bar{W}_T^\pi],$$

In Wang and Zhou (2020), they adopt a constrained formulation of the mean variance problem. However, it is well know that for the Black-Scholes model, two formulations are equivalent (Zhou

Electronic copy available at: https://ssrn.com/abstract=3770818

and Li 2000, Dai et al. 2020). Recall that the pre-committed strategy in Wang and Zhou (2020) is given by

$$\pi^*_{pre-committed} = \mathcal{N}\Big(-\frac{\mu-r}{\sigma^2}(\bar{w}-c), \frac{\lambda}{2\sigma^2}e^{\frac{(\mu-r)^2}{\sigma^2}(T-t)}\Big),$$

where c is a constant depending on model primitives and the risk aversion parameter.

On the other hand, the next proposition shows that the equilibrium policy is given by

$$\pi^*_{equilibrium} = \mathcal{N}\Big(\frac{\mu-r}{\gamma\sigma^2}, \frac{\lambda}{\gamma\sigma^2}\Big). \tag{EC.5}$$

Proposition EC.2. *The control policy in* (EC.5) *is an equilibrium policy.*

Proof of Proposition EC.2 It suffices to verify the definition.

First, under the policy (EC.5), $\bar{W}_t^{\pi^*_{equilibrium}}$ is a Gaussian process, with

$$\mathbb{E}_t\big[\bar{W}_t^{\pi^*_{equilibrium}}\big] = \bar{W} + \frac{(\mu-r)^2}{\gamma\sigma^2}(T-t),\ Var_t\big[\bar{W}_t^{\pi^*_{equilibrium}}\big] = (\frac{(\mu-r)^2}{\gamma^2\sigma^2}+\frac{\lambda}{\gamma})(T-t).$$

If a different strategy π is applied during $[t, t+h)$, then the objective functional becomes

$$\begin{aligned} J(t,\bar{W},\pi^h) =& \mathbb{E}_t\big[\bar{W}^\pi_{t+h} + \frac{(\mu-r)^2}{\gamma\sigma^2}(T-t-h) + \lambda h H(\pi) + \frac{\lambda}{2}(T-t-h)\log\frac{2\pi e\lambda}{\gamma\sigma^2}\big] \\ & - \frac{\gamma}{2}Var_t[\bar{W}^\pi_{t+h}] - \frac{\gamma}{2}(\frac{(\mu-r)^2}{\gamma^2\sigma^2}+\frac{\lambda}{\gamma})(T-t-h) \end{aligned}$$

Therefore,

$$\begin{aligned} & J(t,\bar{W},\pi^h) - J(t,\bar{W},\pi^*_{equilibrium}) \\ =& \mathbb{E}_t\big[\bar{W}^\pi_{t+h} - \bar{W}_t + \lambda h H(\pi)\big] - \frac{\gamma}{2}Var_t[\bar{W}^\pi_{t+h}] - \frac{(\mu-r)^2}{\gamma\sigma^2}h - \frac{\lambda h}{2}\log\frac{2\pi e\lambda}{\gamma\sigma^2} + \frac{\gamma}{2}(\frac{(\mu-r)^2}{\gamma^2\sigma^2}+\frac{\lambda}{\gamma})h \\ =& h[(\mu-r)\int_{\mathbb{R}} u\pi(u)\mathrm{d}u + \lambda H(\pi)] - \frac{\gamma h\sigma^2}{2}\int_{\mathbb{R}} u^2\pi(u)\mathrm{d}u \\ & - \frac{(\mu-r)^2}{\gamma\sigma^2}h - \frac{\lambda h}{2}\log\frac{2\pi e\lambda}{\gamma\sigma^2} + \frac{\gamma}{2}(\frac{(\mu-r)^2}{\gamma^2\sigma^2}+\frac{\lambda}{\gamma})h + o(h) \\ \le& h[(\mu-r)m_1 + \frac{\lambda}{2}\log 2\pi e m_2 - \frac{\gamma\sigma^2}{2}(m_1^2+m_2)] \\ & - h[\frac{(\mu-r)^2}{\gamma\sigma^2} - \frac{\lambda}{2}\log\frac{2\pi e\lambda}{\gamma\sigma^2} + \frac{\gamma}{2}(\frac{(\mu-r)^2}{\gamma^2\sigma^2}+\frac{\lambda}{\gamma})] + o(h) \le o(h), \end{aligned}$$

where m_1, m_2 are introduced to denote the first and the second moment of the distribution π. The last inequality is because as a function of m_1, m_2, it achieves the maximum value at $m_1^* = \frac{\mu-r}{\gamma\sigma^2}$, and $m_2 = \frac{\lambda}{\gamma\sigma^2}$. □

Electronic copy available at: https://ssrn.com/abstract=3770818

We realize that under the same setting (the same model and the same objective function), two solution concepts lead to different solutions. The mean of each policies coincides with its non-exploratory counterpart (Wang and Zhou 2020, Basak and Chabakauri 2010) and the difference in that part has been well understood. While we also identify very different patterns in exploratory variance. The variance of the pre-committed strategy decays in time, on the contrary, the variance of the equilibrium strategy remains a constant. The constant or non-decaying variance also appears in the equilibrium solution under other objective functions (log-MV) and other models (Gaussian mean return model). This finding complements our discussions in Section 3 and it seems that this is a feature of equilibrium solution and provides a new aspect to understand the difference between two solution concepts.

EC.4. Stochastic Volatility Model

In this part, we consider a general stochastic volatility model, where the instantaneous volatility is $\sigma(t,x)=x^{\frac{1}{2\alpha}}$. This term can be taken, for example, from the implied volatility from the option price as a natural proxy. The rest of the model is specified as:

$$\mu(t,x)=r+\delta x^{\frac{1+\alpha}{2\alpha}}, \sigma(t,x)=x^{\frac{1}{2\alpha}}, m(t,x)=\iota(\bar{x}-x) \text{ and } \nu(t,x)=\bar{\nu}\sqrt{x}.$$

In this case, we shall have $\theta(t,x)=\delta x^{\frac{1}{2}}$. Applying Theorem 1, we may prove the following proposition about the theoretical equilibrium policy for this model.

Proposition EC.3. *An equilibrium for the stochastic volatility model is:*

$$\pi^*(t,X,R)\sim\mathcal{N}\Big(\frac{\delta}{1+\gamma}X^{\frac{\alpha-1}{2\alpha}}-\frac{\gamma\rho\bar{\nu}}{(1+\gamma)}a_1(t)X^{\frac{\alpha-1}{2\alpha}}, \frac{\lambda}{(1+\gamma)X^{\frac{1}{\alpha}}}\Big),$$

where

$$a_1(t)=-\frac{(1+2\gamma)\delta^2}{(1+\gamma)^2}\frac{e^{2C_3(T-t)}-1}{(C_3+C_4)(e^{2C_1(T-t)}-1)+2C_3},$$

with $C_3=\frac{1}{1+\gamma}\left[\gamma^2(\iota+\rho\bar{\nu}\delta)^2+\iota^2(2\gamma+1)\right]^{\frac{1}{2}}$ *and* $C_4=\iota+\frac{\gamma^2\rho\delta\bar{\nu}}{(1+\gamma)^2}$.

Moreover, the associated value function has the form of

$$g^*(t,R,X)=R+a_1(t)X+a_2(t),\ V^*(t,R,X)=R+U^*(t,X),$$

Electronic copy available at: https://ssrn.com/abstract=3770818

where $a_0(t)$ is the solution to the ODE $a_0'(t)+\iota\bar{x}a_1(t)-\frac{\lambda}{2(1+\gamma)}+r=0$, with terminal condition $a_0(T)=0$; $U^*(t,X)$ satisfies (10) *but does not admit a closed-form solution.*

Proof of Proposition EC.3 As an application of Theorem 1, it suffices to verify that (8) admits a solution $h^*(t,X)=a_1(t)X+a_0(t)$. It can be done by direct calculation. □

EC.5. The PDE Characterization of Value Functions

To proceed with the PDE approach, we shall have the following verification theorem under the following regularity condition.

ASSUMPTION EC.1. *(i) $\mu(\cdot,\cdot)$, $\sigma(\cdot,\cdot)$, $m(\cdot,\cdot)$ and $\nu(\cdot,\cdot)$ are continuous functions;*

(ii) μ,σ are polynomial growth with respect to X and m,ν are linear growth with respect to x, i.e. there exist positive constants C and p such that

$$|\mu(t,X)|,|\sigma(t,X)|\le C(1+|X|^p),$$

and

$$|m(t,X)|,|\nu(t,X)|\le C(1+|X|).$$

THEOREM EC.1. *Suppose $V^*(t,R,X)=R+U^*(t,X)$ and $g^*(t,R,X)=R+h^*(t,X)$ are the classical solutions to the extended HJB system* (14) *and* (15)*., with the equilibrium condition* (17)*, where If U^*, g^* and their derivatives have polynomial growth with respect to x, and Assumption EC.1 holds true.*

Then the feedback strategy π^* defined by

$$\pi^*(t,X)\sim\mathcal{N}\Big(\frac{\mu(t,X)-r-\rho\gamma\sigma(t,X)\nu(t,X)\partial_X h^*(t,X)}{(1+\gamma)\sigma^2(t,X)},\frac{\lambda}{(1+\gamma)\sigma^2(t,X)}\Big),\tag{EC.6}$$

is an equilibrium strategy. In particular h^* solves (8) *and* (10)*.*

Proof of Theorem EC.1 We make the ansatz that $V^*(t,R,X)=R+U^*(t,X)$ and $g^*(t,R,X)=R+h^*(t,X)$. Then, the above equations can be further simplified as

$$\begin{aligned}&\partial_t U^*+m(t,x)\partial_X U^*+\frac{1}{2}\nu^2(t,X)\partial_{XX}U^*+\left[r+(\mu(t,X)-r)\int_{\mathbb{R}}u\pi^*(du)-\frac{1}{2}\sigma^2(t,X)\int_{\mathbb{R}}u^2\pi^*(du)\right]\\&\quad-\gamma\left[\frac{1}{2}\sigma^2(t,X)\int_{\mathbb{R}}u^2\pi^*(du)+\rho\nu(t,X)\sigma(t,X)\partial_X g\int_{\mathbb{R}}u\pi^*(du)+\frac{1}{2}\nu^2(t,x)|\partial_X g^*|^2\right]+\lambda H(\pi^*)=0,\end{aligned}\tag{EC.7}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

$$\partial_t h^* + m(t,x)\partial_X h^* + \frac{1}{2}\nu^2(t,X)\partial_{XX} h^* + \left[r + (\mu(t,X) - r)\int_{\mathbb{R}} u\pi^*(du) - \frac{1}{2}\sigma^2(t,X)\int_{\mathbb{R}} u^2\pi^*(du)\right] = 0, \tag{EC.8}$$

with $U^*(T,X) = 0$, $h^*(T,X) = 0$, and

$$\pi^*(t,x) = \underset{\pi\in\mathcal{P}(\mathbb{R})}{\operatorname{argmax}} \Bigg\{ (\mu(t,X) - r)\int_{\mathbb{R}} u\pi(du) - \frac{1}{2}\sigma^2(t,X)\int_{\mathbb{R}} u^2\pi(du) \\ - \gamma\left[\frac{1}{2}\sigma^2(t,X)\int_{\mathbb{R}} u^2\pi(du) + \rho\nu(t,X)\sigma(t,X)\partial_X h^* \int_{\mathbb{R}} u\pi(du)\right] + \lambda H(\pi)\Bigg\}. \tag{EC.9}$$

Note that, on the right hand side of (EC.9), except the entropy term, other terms only depend on π through the first and second moment $\int_{\mathbb{R}} u\pi(du)$ and $\int_{\mathbb{R}} u^2\pi(du)$. From Cover and Thomas (2012), we know that, of all the probability distributions over the reals with a specified mean and variance, normal distribution is the one with the maximal entropy. Hence, π^* should be a normal distribution. Choosing its mean and variance to maximize the right hand side of (EC.9), we have that (EC.6).

The proof consists of two steps:

- We first prove that $V^*(t,R,X) := R + U^*(t,X)$ and $g^*(T,R,X) := R + h^*(t,X)$ are the desired functions, i.e. $V^*(t,R,X) = J(t,R,X;\pi^*)$ and $g^*(t,R,X) = \mathbb{E}_t\left[R_T^{\pi^*} | R_t = R, X_t = X\right]$;
- In the second step, we show that π^* is the equilibrium strategy.

First note that, since μ, m are linear growth with respect to x, we have, for any $p > 1$, there exists a constant C_p such that

$$\mathbb{E}\left[\sup_{0\le t\le T} |X_t|^p\right] \le C_p |X_0|^p.$$

From (8), applying Itô formula to $g^*(t, R_t^{\pi^*}, X_t)$, we have

$$dg^*(t, R_t^{\pi^*}, X_t) = (\partial_t + \mathcal{A}_t^{\pi^*})g^* dt + \nu\partial_x g^*[\rho dB_s + \sqrt{1-\rho^2}d\tilde{B}_s] \\ + \sigma(s, X_s)\left[\int u\pi_t^*(du)dB_s + \sqrt{\int_{\mathbb{R}} u^2\pi_s^*(du) - (\int_{\mathbb{R}} u\pi_s^*(du))^2}\, d\bar{B}_s\right]$$

Since g^* satisfying the extended HJB equation, the dt term on the right hand side of above equation is identical to zero. Moreover, from the growth condition on the coefficients and g^*, it follows that

Electronic copy available at: https://ssrn.com/abstract=3770818

$g^*(t, R_t^{\pi^*}, X_t)$ is a martingale. So, by the boundary condition of g^*, it is the expectation function of π^*. Combining (14) and (15), we have that

$$(\partial_t + \mathcal{A}_t^{\pi^*})V^* - \frac{\gamma}{2}(\partial_t + \mathcal{A}_t^{\pi^*})g^2 + \lambda H(\pi_t^*) = 0.$$

Using Itô formula and the boundary condition of V^*, we have

$$\begin{aligned}
V^*(t, R_t, X_t) =& \mathbb{E}_t\left[R_T^{\pi^*} + \lambda\int_t^T H(\pi_s^*)ds\right] - \frac{\gamma}{2}\mathbb{E}_t\left[\int_t^T (\partial_s + \mathcal{A}_s^{\pi^*})(g^*)^2 ds\right] \\
=& \mathbb{E}_t\left[R_T^{\pi^*} + \lambda\int_t^T H(\pi_s^*)ds\right] - \frac{\gamma}{2}((g^*)^2(T, R_T^{\pi^*}, X_T) - (g^*)^2(t, R_t, X_t)) \\
=& \mathbb{E}_t\left[R_T^{\pi^*} + \lambda\int_t^T H(\pi_s^*)ds\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_T^{\pi^*}\right],
\end{aligned}$$

where the last equality is obtained due the fact that g^* is the expectation of the terminal log-return.

Now we are going to show that π^* is an equilibrium strategy. At time t, given any $h \in \mathbb{R}^+$ and $v \in \mathcal{P}(\mathbb{R})$, consider the perturbation strategy $\pi^{h,v}$ as defined in Definition 2. Note that

$$\mathrm{Var}_t\left[R_T^{\pi^{h,v}}\right] = \mathbb{E}_t\left[\mathrm{Var}_{t+h}\left[R_T^{\pi^{h,v}}\right]\right] + \mathrm{Var}_t\left[\mathbb{E}_{t+h}\left[R_T^{\pi^{h,v}}\right]\right]$$

Since $\pi_s^{h,v} = \pi_s^*$ for $s \in [t+h, T]$, we have $\mathrm{Var}_{t+h}\left[R_T^{\pi^{h,v}}\right] = \mathrm{Var}_{t+h}\left[R_T^{\pi^*}\right]$ and $\mathbb{E}_{t+h}\left[R_T^{\pi^{h,v}}\right] = \mathbb{E}_{t+h}\left[R_T^{\pi^*}\right] = g(t+h, R_{t+h}^{\pi^{h,v}}, X_{t+h})$. Thus,

$$\begin{aligned}
J(t, R_t, X_t; \pi^{h,v}) =& \mathbb{E}_t\left[R_T^{\pi^{h,v}} + \lambda\int_t^T H(\pi_s^{h,v})ds\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_T^{\pi^{h,v}}\right] \\
=& \mathbb{E}_t\left[\lambda\int_t^{t+h} H(\pi_s^{h,v})ds\right] + \mathbb{E}_t\left[\mathbb{E}_{t+h}\left[R_T^{\pi^{h,v}} + \lambda\int_{t+h}^T H(\pi_s^{h,v})ds\right]\right] \\
& - \frac{\gamma}{2}\left(\mathbb{E}_t\left[\mathrm{Var}_{t+h}\left[R_T^{\pi^{h,v}}\right]\right] + \mathrm{Var}_t\left[\mathbb{E}_{t+h}\left[R_T^{\pi^{h,v}}\right]\right]\right) \\
=& \mathbb{E}_t\left[V^*(t+h, R_{t+h}^{\pi^{h,v}}, X_{t+h}) + \lambda\int_t^{t+h} H(\pi_s^{h,v})ds\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[g^*(t+h, R_{t+h}^{\pi^{h,v}}, X_{t+h})\right].
\end{aligned}$$

Applying Itô formula to $V^*(s, R_s^{\pi^{h,v}}, X_s)$, we see that

$$\mathbb{E}_t\left[V^*(t+h, R_{t+h}^{\pi^{h,v}}, X_{t+h}) + \lambda\int_t^{t+h} H(\pi_s^{h,v})ds\right] = V^*(t, R_t, X_t) + \mathbb{E}_t\left[\int_t^{t+h} (\partial_s + \mathcal{A}^v)V^* + \lambda H(v)ds\right]$$

From (14) and the optimality condition (17) of π, we have

$$\begin{aligned}
& \mathbb{E}_t\left[V^*(t+h, R_{t+h}^{\pi^{h,v}}, X_{t+h}) + \lambda\int_t^{t+h} H(\pi_s^{h,v})ds\right] \\
\leq& V^*(t, R_t, X_t) - \frac{\gamma}{2}\mathbb{E}_t\left[\int_t^{t+h} (2g^*\mathcal{A}^v g^* - \mathcal{A}^v(g^*)^2)(s, R_s^{\pi^{h,v}}, X_s)ds\right] \\
=& V^*(t, R_t, X_t) - \frac{\gamma}{2}\mathbb{E}_t\left[\int_t^{t+h} (2g^*\partial_s g^* + 2g^*\mathcal{A}^v g^* - \partial_s(g^*)^2 - \mathcal{A}^u(g^*)^2)(s, R_s^{\pi^{h,v}}, X_s)ds\right].
\end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

Furthermore, it holds that

$$\mathbb{E}_t\left[\int_t^{t+h}(\partial_s(g^*)^2+\mathcal{A}^v(g^*)^2)(s,R_s^{\pi^{h,v}},X_s)ds\right]=\mathbb{E}_t\left[(g^*)^2(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})\right]-(g^*)^2(t,R_t,X_t)$$

and

$$\begin{aligned}&\mathbb{E}_t\left[\int_t^{t+h}(g^*\partial_s g^*+g^*\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]\\=&g^*(t,R_t,X_t)\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]+I\end{aligned}$$

with

$$I=\mathbb{E}\left[\int_t^{t+h}(g^*(s,R_s,X_s)-g^*(t,R_s,X_s))(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right].$$

Hölder inequality yields that

$$|I|^2\le\left(\mathbb{E}\left[\int_t^{t+h}(g^*(s,R_s,X_s)-g^*(t,R_s,X_s))^2ds\right]\right)^{1/2}\left(\mathbb{E}\left[\int_t^{t+h}((\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s))^2ds\right]\right)^{1/2}.$$

It further implies that $I=o(h)$. Then, we have

$$\begin{aligned}&\mathbb{E}_t\left[V^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})+\lambda\int_t^{t+h}H(\pi_s^{h,v})ds\right]\\\le&V^*(t,R_t,X_t)+\frac{\gamma}{2}\Bigg(\mathbb{E}_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})^2\right]-(g^*)^2(t,R_t,X_t)\\&-2g^*(t,R_t,X_t)\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]\Bigg)+o(h)\end{aligned}\tag{EC.10}$$

On the other hand,

$$\mathrm{Var}_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})\right]=\mathbb{E}_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})^2\right]-\left(\mathbb{E}_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})\right]\right)^2.$$

We also have

$$E_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})\right]=g^*(t,R_t,X_t)+\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right].$$

Thus,

$$\begin{aligned}&\left(\mathbb{E}_t\left[g^*(t+h,R_{t+h}^{\pi^{h,v}},X_{t+h})\right]\right)^2\\=&(g^*)^2(t,R_t,X_t)+2g^*(t,R_t,X_t)\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]\\&+\left(\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]\right)^2\\=&(g^*)^2(t,R_t,X_t)+2g^*(t,R_t,X_t)\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R_s^{\pi^{h,v}},X_s)ds\right]+o(h).\end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

This implies that

$$\begin{aligned}\mathrm{Var}_t\left[g^*(t+h,R^{\pi^{h,v}}_{t+h},X_{t+h})\right]=&\mathbb{E}_t\left[g^*(t+h,R^{\pi^{h,v}}_{t+h},X_{t+h})^2\right]-(g^*)^2(t,R_t,X_t)\\&-2g^*(t,R_t,X_t)\mathbb{E}_t\left[\int_t^{t+h}(\partial_s g^*+\mathcal{A}^v g^*)(s,R^{\pi^{h,v}}_s,X_s)ds\right]+o(h)\end{aligned}$$

Combining this with (EC.10), we shall have

$$J(t,R_t,X_t;\pi^{h,v})\le V^*(t,R_t,X_t)+o(h),$$

which indicates that π^* is an equilibrium strategy. □

In addition, as a side product of Theorem EC.1, we have the following corollary that characterizes the value function and expectation function under any strategy.

COROLLARY EC.1. *Under Assumption EC.1, for any given admissible strategy π, the related value function V^π and the expectation function g^π satisfy PDE system* (14) *and* (15).

The proof of Corollary EC.1 is parallel to one step in the proof of Theorem EC.1 and hence is omitted. We would like to emphasis that this result is crucial for the design of the algorithm, as we need to evaluate any policy, which is equivalent to solve (14). and (15) for any given strategy π.

EC.6. Proof of Statements

EC.6.1. Proof of Theorem 1

Proof of Theorem 1 (i) From the assumption of the theorem, we see that BSDE (7) admits a unique solution (Y,Z). Consider the strategy π^* defined by (6). One can verify that it is admissible. The related log-return process is denoted as R^*. Direct computation shows that

$$Y_t=\mathbb{E}_t\left[\int_t^T f(s,X_s,Z_s)ds\right]=\mathbb{E}\left[R^*_T-R^*_t\right].$$

Electronic copy available at: https://ssrn.com/abstract=3770818

For any local perturbation, $\pi_s^{h,v} = \pi_s^* 1_{s\notin[t,t+h)} + v1_{s\in[t,t+h)}$, denote the related log-return process by $R^{h,v}$. Then, $R_T^{h,v} - R_{t+h}^{h,v} = R_T^* - R_{t+h}^*$, which implies that $\mathbb{E}_{t+h}\left[R_T^{h,v}\right] = R_{t+h}^{h,v} + Y_{t+h}$. It follows that

$$
\begin{aligned}
&J(t, X_t, R_t; \pi^{h,v}) \\
=&E_t\left[R_T^{h,v} + \int_t^T \lambda H(\pi^{h,v})\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_T^{h,v}\right] \\
=&E_t\left[\mathbb{E}_{t+h}\left[R_T^{h,v}\right] - \frac{\gamma}{2}\mathrm{Var}_{t+h}\left[R_T^{h,v}\right]\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[E_{t+h}\left[R_T^{h,v}\right]\right] + E_t\left[\int_t^T \lambda H(\pi_s^{h,v})ds\right] \\
=&E_t\left[R_{t+h}^{h,v} + Y_{t+h} - \frac{\gamma}{2}\mathrm{Var}_{t+h}\left[R_T^{h,v} - R_{t+h}^{h,v}\right]\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_{t+h}^{h,v} + Y_{t+h}\right] + E_t\left[\int_t^T \lambda H(\pi_s^{h,v})ds\right] \\
=&E_t\left[R_{t+h}^{*} + Y_{t+h} - \frac{\gamma}{2}\mathrm{Var}_{t+h}\left[R_T^{*} - R_{t+h}^{*}\right]\right] - \frac{\gamma}{2}\mathrm{Var}_t\left[R_{t+h}^{h,v} + Y_{t+h}\right] \\
&+ \mathbb{E}_t\left[R_{t+h}^{h,v} - R_{t+h}^{*}\right] + E_t\left[\int_t^T \lambda H(\pi_s^{h,v})ds\right] \\
=&J(t, X_t, R_t; \pi^*) + \mathbb{E}_t\left[R_{t+h}^{h,v} - R_{t+h}^{*}\right] + E_t\left[\int_t^{t+h} \lambda H(v) - \lambda H(\pi_s^*)ds\right] \\
&- \frac{\gamma}{2}\mathrm{Var}_t\left[R_{t+h}^{h,v} + Y_{t+h}\right] + \frac{\gamma}{2}\mathrm{Var}_t\left[R_{t+h}^{*} + Y_{t+h}\right].
\end{aligned}
$$

Note that

$$
R_{t+h}^{h,v} = R_t + \int_t^{t+h} a(s,v)ds + \int_t^{t+h} \sigma_s\left[\int uv(du)dB_s + \sqrt{\int u^2 v(du) - (\int uv(du))^2}d\bar{B}_s\right],
$$

with

$$
a(s,v) = r + (\mu(s,X_s) - r)\int uv(du) - \frac{1}{2}\sigma^2(s,X_s)\int u^2 v(du).
$$

Thus, we have

$$
\mathbb{E}_t\left[R_{t+h}^{h,v} - R_{t+h}^{*}\right] = \mathbb{E}_t\left[\int_t^{t+h} a(s,v) - a(s,\pi_s^*)ds\right]
$$

It is not hard to compute that

$$
\mathrm{Var}_t\left[R_{t+h}^{h,v} + Y_{t+h}\right] = \mathbb{E}_t\left[\int_t^{t+h} \phi(s,v)ds\right] + o(h),
$$

and

$$
\mathrm{Var}_t\left[R_{t+h}^{*} + Y_{t+h}\right] = \mathbb{E}_t\left[\int_t^{t+h} \phi(s,\pi_s^*)ds\right] + o(h),
$$

where

$$
\phi(s,v) = Z_s^2 + \sigma(s,X_s)\int u^2 v(du) + 2\rho\sigma(s,X_s)Z_s\int uv(du).
$$

Electronic copy available at: https://ssrn.com/abstract=3770818

Hence,

$$\lim_{h\to 0}\frac{J(t,X_t,R_t;\pi^{h,v})-J(t,X_t,R_t;\pi^*)}{h}=a(s,v)+\phi(s,v)+\lambda H(v)-\left(a(s,\pi^*)+\phi(s,\pi^*)+\lambda H(\pi^*)\right).$$

Finally, we observe that the random functional $a(s,\cdot)+\phi(s,\cdot)+\lambda H(\cdot)$ is minimized at π_s^*. This implies that π^* is the equilibrium strategy.

(ii) In this case, the BSDE (7) is Markovian. Hence, we can write $h=f(t,X)$ for a deterministic function h. If $h\in C^{1,2}$, we can apply Itô formula to get the desired result.

(iii) Based on the previous argument, we could get that $R_T^{\pi^*}=R_t^{\pi^*}+\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s$. Therefore

$$g^*(t,R,X)=\mathbb{E}_t[R_T^{\pi^*}|R_t^{\pi^*}=R,X_t=X]=R+\mathbb{E}_t[\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s]=R+h^*(t,X),$$

and

$$\begin{aligned}V^*(t,R,X)=&\mathbb{E}_t[R_t^{\pi^*}+\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s+\lambda H(\pi_s^*)\mathrm{d}s|R_t^{\pi^*}=R,X_t=X]\\&-\frac{\gamma}{2}Var_t[R_t^{\pi^*}+\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s|R_t^{\pi^*}=R,X_t=X]\\=&R+\mathbb{E}_t[\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s+\lambda H(\pi_s^*)\mathrm{d}s]\\&-\frac{\gamma}{2}Var_t[\int_t^T\alpha(s,X_s)\mathrm{d}s+\eta(s,X_s)\mathrm{d}B_s]\\=&R+U^*(t,X),\end{aligned}$$

for some function h^* and U^* that only depend on t and X. By taking $t=T$, we can obtain that they satisfy the terminal condition $h^*(T,X)=U^*(T,X)=0$.

Note that it means $R_t^{\pi^*}+h^*(t,X_t)$ is a martingale, under the equilibrium policy π^*. If $h^*\in C^{1,2}$, we can apply Itô's lemma to deduce that h^* satisfies PDE (8).

Moreover, we can rewrite the above identity as

$$\begin{aligned}R+U^*(t,X)=&\mathbb{E}_t[R_T^{\pi^*}+\int_t^T\lambda H(\pi_s)\mathrm{d}s]-\frac{\gamma}{2}Var_t[R_T^{\pi^*}]\\=&\mathbb{E}_t[R_T^{\pi^*}+U^*(T,X_T)+\int_t^T\frac{\lambda}{2}\log 2\pi e\frac{\lambda}{(1+\gamma)\sigma^2(s,X_s)}\mathrm{d}s]-\frac{\gamma}{2}Var_t[R_T^{\pi^*}+h^*(T,X_T)]\\=&\mathbb{E}_t[R_T^{\pi^*}+U^*(T,X_T)+\int_t^T\frac{\lambda}{2}\log 2\pi e\frac{\lambda}{(1+\gamma)\sigma^2(s,X_s)}\mathrm{d}s-\frac{\gamma}{2}\mathrm{d}\langle R^{\pi^*}+h^*\rangle(s)].\end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

It means that $R_t^{\pi^*}+U^*(t,X_t)+\int_0^t\frac{\lambda}{2}\log 2\pi e\frac{\lambda}{(1+\gamma)\sigma^2(s,X_s)}\mathrm{d}s-\frac{\gamma}{2}\langle R+h^*\rangle(t)$ is a martingale. If $U^*\in C^{1,2}$, then we can apply Itô's lemma to deduce that U^* satisfies PDE (10). $\square$

EC.6.2. Proof of Proposition 1

Proof of Proposition 1 Applying Theorem 1, we shall have $\theta(t,X)=X$ and h^*,U^* satisfy

$$\begin{aligned}&\partial_t h^*+\frac{\nu^2}{2}\partial_{XX}h^*+\iota(\bar{X}-X)\partial_X h^*+r-\frac{\lambda}{2(1+\gamma)}+\frac{1}{2}X^2-\frac{\gamma^2}{2(1+\gamma)^2}(X+\rho\nu\partial_X h^*)^2=0,\\&h^*(T,X)=0,\end{aligned}\tag{EC.11}$$

and

$$\begin{aligned}&\partial_t U^*+\frac{\nu^2}{2}\partial_{XX}U^*+\iota(\bar{X}-X)\partial_X U^*+r-\frac{1}{2}\frac{(X-\gamma\rho\nu\partial_X h^*)^2}{(1+\gamma)\sigma}-\frac{1}{2}\nu^2(\partial_x h^*)^2+\frac{\lambda}{2}\left[\log\frac{2\pi e\lambda}{(1+\gamma)\sigma^2}-1\right]=0,\\&U^*(T,X)=0.\end{aligned}\tag{EC.12}$$

In this case, $h^*(t,X)$ can be solved as $h^*(t,X)=\frac{1}{2}a_2^*(t)X^2+a_1^*(t)X+a_0^*(t)$ with $a_i^*,i=1,2,3$, satisfying the following ODEs

$$\begin{cases}\dot{a}_2^*(t)=\dfrac{\gamma^2\rho^2\nu^2}{(1+\gamma)^2}(a_2^*)^2(t)+2(\iota+\dfrac{\gamma^2\rho\nu}{(1+\gamma)^2})a_2^*(t)-\dfrac{1+2\gamma}{(1+\gamma)^2},a_2^*(T)=0,\\ \dot{a}_1^*(t)=(\iota+\dfrac{\gamma^2\rho\nu}{(1+\gamma)^2}+\dfrac{\gamma^2\rho^2\nu^2}{(1+\gamma)^2}a_2^*(t))a_1^*(t)-\iota\bar{X}a_2^*(t),a_1^*(T)=0,\\ \dot{a}_0^*(t)=\dfrac{\gamma^2\rho^2\nu^2}{2(1+\gamma)^2}(a_1^*)^2(t)-\iota\bar{X}a_1^*(t)-\dfrac{1}{2}\nu^2a_2^*(t)-(r-\dfrac{\lambda}{2(1+\gamma)}),a_0^*(T)=0.\end{cases}\tag{EC.13}$$

We can explicitly solve these equations. Especially, we have that

$$\begin{cases}a_2^*(t)=\dfrac{(1+2\gamma)}{(1+\gamma)^2}\dfrac{e^{2C_1(T-t)}-1}{(C_1+C_2)(e^{2C_1(T-t)}-1)+2C_1},\\ a_1^*(t)=\dfrac{\iota\bar{X}(1+2\gamma)}{(1+\gamma)^2}\dfrac{(e^{C_1(T-t)}-1)^2}{C_1[(C_1+C_2)(e^{2C_1(T-t)}-1)+2C_1]}.\end{cases}$$

where $C_1=\frac{1}{\gamma+1}\left[\gamma^2(\iota+\rho\nu)^2+\iota^2(2\gamma+1)\right]^{\frac{1}{2}}$,$C_2=\iota+\frac{\gamma^2\rho\nu}{(1+\gamma)^2}$. It is exactly the same function obtained in Dai et al. (2020). Similarly, $U^*(t,X)$ can be also written as $U^*(t,X)=\frac{1}{2}b_2^*(t)X^2+b_1^*(t)X+b_0^*(t)$

Electronic copy available at: https://ssrn.com/abstract=3770818

with b_i^* satisfying certain ODEs:

$$\begin{cases} \dot{b}_2^*(t) = 2\iota b_2^*(t) + \dfrac{1}{2(1+\gamma)}(\gamma\rho\nu a_2^*(t) - 1)^2 + \dfrac{\nu^2}{2}(a_2^*)^2(t), b_2^*(T) = 0, \\ \dot{b}_1^*(t) = \iota b_1^*(t) - \iota\bar{X} b_2^*(t) - \dfrac{\rho\nu}{1+\gamma}(1+\gamma\rho\nu a_2^*(t))a_1^*(t) + \nu^2 a_1^*(t)a_2^*(t), b_1^*(T) = 0, \\ \dot{b}_0^*(t) = -\dfrac{\nu^2}{2}b_2^*(t) - \iota\bar{X} b_1^*(t) - r + \dfrac{\gamma^2\rho^2\nu^2(a_1^*)^2(t)}{2(1+\gamma)} + \dfrac{\nu^2}{2}a_1^*(t) + \dfrac{\lambda}{2}\left[\log\dfrac{2\pi e\lambda}{(1+\gamma)\sigma^2} - 1\right], b_0^*(T) = 0. \end{cases} \tag{EC.14}$$

□

EC.6.3. Proof of Theorem 2

Proof of Proposition 2 We know that the related value function $V^{(0)}(t,R,X) := J(t,R,X;\pi^{(0)})$ and the expectation function $g^{(0)}(t,R,X) := \mathbb{E}\left[R_T^{\pi^{(0)}} | R_t = R, X_t = X\right]$ can be written as $V^{(0)}(t,R,X) = R + U^{(0)}(t,X)$ and $g^{(0)}(t,R,X) = R + h^{(0)}(t,X)$ with $V^{(0)}$ and $h^{(0)}$ satisfying

$$\begin{aligned} \partial_t h^{(0)} + \frac{\nu^2}{2}\partial_{XX}h^{(0)} + \iota(\bar{X} - X)\partial_X h^{(0)} + r - \frac{\lambda}{2(1+\gamma)} + \frac{1}{2}X^2 \\ - \frac{\gamma^2}{2(1+\gamma)^2}(X + \rho\nu[a_2^{(0)}(t)X + a_1^{(0)}(t)])^2 = 0, h^{(0)}(T,X) = 0, \end{aligned}$$

and

$$\begin{aligned} \partial_t U^{(0)} + \frac{\nu^2}{2}\partial_{XX}U^{(0)} + \iota(\bar{X} - X)\partial_X U^{(0)} + r - \frac{1}{2}\frac{(X - \gamma\rho\nu[a_2^{(0)}(t)X + a_1^{(0)}(t)])^2}{(1+\gamma)} \\ - \gamma\rho\nu(\partial_X h^{(0)} - (a_2^{(0)}(t)X + a_1^{(0)}(t)))(a_2^{(0)}(t)X + a_1^{(0)}(t)) - \frac{1}{2}\nu^2(\partial_x h^{(0)})^2 \\ + \frac{\lambda}{2}\log 2\pi e\theta^{(0)} - \frac{1+\gamma}{2}\sigma^2\theta^{(0)} = 0, U^{(0)}(T,X) = 0. \end{aligned}$$

Furthermore, $U^{(0)}$ and $h^{(0)}$ can be solved as

$$h^{(0)}(t,X) = \frac{1}{2}a_2^{(1)}(t)X^2 + a_1^{(1)}(t)X^1 + a^{(0)}(t)$$

and

$$U^{(0)}(t,X) = \frac{1}{2}b_2^{(1)}(t)X^2 + b_1^{(1)}(t)X^1 + b^{(0)}(t)$$

with $a_i^{(1)}$ and $b_i^{(1)}$ satisfying ODEs

$$\begin{cases} \dot{a}_2^{(1)} = 2\iota a_2^{(1)} + \dfrac{\gamma^2}{(1+\gamma)^2}(1 + \rho\nu a_2^{(0)}(t))^2 - 1, a_2^{(1)}(T) = 0, \\ \dot{a}_1^{(1)} = \iota a_1^{(1)} - \iota a_2^{(1)} + \dfrac{\gamma^2\rho\nu}{(1+\gamma)^2}(1 + \rho\nu a_2^{(0)})a_1^{(0)}, a_1^{(1)}(T) = 0, \\ \dot{a}_0^{(1)} = -\iota\bar{X}a_1^{(1)} - \dfrac{\nu^2}{2}a_2^{(1)} - (r - \dfrac{\lambda}{2(1+\gamma)}) + \dfrac{\gamma^2\rho^2\nu^2}{2(1+\gamma)^2}a_1^{(0)}, a_0^{(1)}(T) = 0, \end{cases}$$

Electronic copy available at: https://ssrn.com/abstract=3770818

and

$$\begin{cases} \dot{b}_2^{(1)} =& 2\iota b_2^{(1)} + \dfrac{1}{1+\gamma}(\gamma\rho\nu a_2^{(0)} - 1)^2 - \gamma\rho\nu(a_2^{(1)} - a_2^{(0)})a_2^{(1)} + \dfrac{\nu^2}{2}a_2^{(1)}, b_2^{(1)}(T) = 0, \\ \dot{b}_1^{(1)} =& \iota b_1^{(1)} - \iota\bar{X} b_2^{(1)} - \dfrac{\rho\nu}{1+\gamma}(1 - \gamma\rho\nu a_2^0)a_1^{(0)} \\ & + \gamma\rho\nu\left[(a_1^{(1)} - a_1^{(0)})a_2^{(1)} + (a_2^{(1)} - a_2^{(0)})a_1^{(1)}\right] + \nu^2 a_2^{(1)} a_1^{(1)}, b_1^{(1)}(T) = 0, \\ \dot{b}_0^{(1)} =& -\iota\bar{X} b_1^{(1)} - \dfrac{\nu^2}{2}b_2^{(1)} - r + \dfrac{\gamma^2\rho^2\nu^2(a_1^{(0)})^2}{2(1+\gamma)^2} + \gamma\rho\nu(a_1^{(1)} - a_1^{(0)})a_1^{(1)} \\ & + \dfrac{\nu^2}{2}(a_1^{(1)})^2 - \left[\dfrac{\lambda}{2}\log 2\pi e\theta^{(0)} - \dfrac{1+\gamma}{2}\sigma^2\theta^{(0)}\right], b_0^{(1)}(T) = 0. \end{cases}$$

Next, we update the policy according to optimality condition (17) with V^*, g^* replaced by $V^{(0)}$ and $g^{(0)}$. Then, the obtained strategy $\pi^{(1)}$ is

$$\pi^{(1)}(t,X) \sim N(\frac{X}{(1+\gamma)\sigma} - \frac{\gamma\rho\nu}{(1+\gamma)\sigma}(a_2^{(1)}(t)X + a_1^{(1)}(t)), \frac{\lambda}{(1+\gamma)\sigma^2}).$$

We repeat this procedure and obtain a sequence of strategies $\{\pi^{(n)}\}$ represented as

$$\pi^{(n)}(t,X) \sim N(\frac{X}{(1+\gamma)\sigma} - \frac{\gamma\rho\nu}{(1+\gamma)\sigma}(a_2^{(n)}(t)X + a_1^{(n)}(t)), \frac{\lambda}{(1+\gamma)\sigma^2}),$$

where $a^{(n)}, i = 1, 2$ satisfying (19).

Next, we move on to prove the convergence results.

(i) Without loss of generality, we may assume that $T = 1$. If it is not the case, consider the function $\tilde{a}_t = a_{tT}$. Let us prove the convergence of $a^{2,(n)}$ first. Then, the convergence of $a^{(1)}$ is an immediate consequence. Denote by $M := \|a_2^*\|$ and assume that $\|a_2^{(0)} - a_2^*\| \le \varepsilon$. Define a sequence $\{K_n\}_n$ as

$$K_0 = \varepsilon, K_{n+1} = \frac{A}{n+1}K_n + \frac{B}{n+1}K_n^2$$

with $A = e^{2\iota}\frac{\gamma^2}{1+\gamma^2}(2 + 2\rho\nu M)$ and $B = e^{2\iota}\frac{\gamma^2}{1+\gamma^2}\rho\nu$. We claim that

$$|a_2^{(n)}(t) - a_2^*(t)| \le K_n(1-t)^n. \tag{EC.15}$$

We prove (EC.15) by induction. It clearly holds when $n = 0$. Assume that it holds for $n = k$. Denote by $\delta_k = a^{(k)} - a^*$. Then, we see that δ_k satisfies the following ODE:

$$\dot{\delta}_{k+1} = 2\iota\delta_{k+1} + \frac{\gamma^2}{1+\gamma^2}(2 + \rho\nu a_2^{(k)} + \rho\nu a^*)(a^{(k)} - a^*)$$

Electronic copy available at: https://ssrn.com/abstract=3770818

Thus, δ_k can be explicitly solved as

$$\delta_{k+1}(t) = -\int_t^1 e^{2\iota(s-t)}\frac{\gamma^2}{1+\gamma^2}(2+\rho\nu a_2^{(k)}+\rho\nu a^*)(a^{(k)}-a^*)ds$$

Let $C_1 = e^{2\iota}\frac{\gamma^2}{1+\gamma^2}$. Then, we have

$$|\delta_{k+1}(t)| \le C_1\int_t^1 (2+2\rho\nu M+\rho\nu|\delta_k(s)|)|\delta_k(s)|ds$$

From the claim that $|\delta_k(s)| \le K_k(1-s)^n$, we have that

$$\begin{aligned}|\delta_{k+1}(t)| \le& C_1\int_t^1 (2+2\rho\nu M+\rho\nu K_k(1-s)^k)K_k(1-s)^k\\ =&\frac{C_1(2+2\rho\nu M)K_k}{k+1}(1-s)^{k+1}+\frac{C_1\rho\nu K_k^2}{2k+1}(1-s)^{2k+1}\\ \le&\left[\frac{C_1(2+2\rho\nu M)}{k+1}K_k+\frac{C_1\rho\nu}{k+1}K_k^2\right](1-s)^{k+1}.\end{aligned}$$

Hence, we prove the claim. By the definition of K_n, we see that

$$\frac{K_{n+1}}{K_n} = \frac{A+BK_n}{n+1}.$$

Thus, we have that

$$K_{n+1} = \frac{\varepsilon}{(n+1)!}\prod_{i=0}^{n}(A+BK_i).$$

We choose $\varepsilon < 1$ such that $\varepsilon \sup_n \frac{(A+B)^{n+1}}{(n+1)!} < 1$. Then, it is easy to get that $K_n \le 1$ and $K_n \le \varepsilon\frac{(A+B)^n}{n!}$.

This also implies that

$$|a_2^{(n)}(t) - a_2^*(t)| \le \varepsilon\frac{(A+B)^n}{n!}(1-t)^n.$$

Thus, we obtain the convergence of $a_2^{(n)}$.

(ii) Define $B(M,T)$ as

$$B(M,T) := \left\{a\in C[0,T]\middle| a(t)\in[-M,T] \text{ for any } t\in[0,T]\right\}.$$

By definition, a^* satisfies

$$\dot{a}_2^* = 2\iota a_2^* + \frac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu a_2^*(t))^2 - 1 \ge 2\iota a_2^* - 1.$$

Electronic copy available at: https://ssrn.com/abstract=3770818

Applying comparison principle of ODEs, it holds that

$$a_2^*(t) \leq \frac{1-e^{-2\iota(T-t)}}{2\iota} \leq T,$$

where we use the inequality $e^{-x}+x-1 \geq 0$, for all $x \geq 0$, to obtain the last inequality. On the other hand, we have $C_1 \geq |C_2|$, which implies that $a_2^*(t) \geq 0$. Thus, we see that $a^* \in B(M,T)$. Now let us show that $a_2^{(n)} \in B(M,T)$ for $n = 0,1,2,.....$ We prove by induction. By the assumption of the theorem, it holds for $a_2^{(0)}$ obviously. Assume that it also holds for $a_2^{(n)}$. Then,

$$2\iota a_2^{(n+1)}(t) - 1 \leq 2\iota a_2^{(n+1)}(t) + \frac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu a_2^{(n)}(t))^2 - 1 \leq 2\iota a_2^{(n+1)}(t) + C(M,T).$$

Using comparison principle agian, we have

$$-C(M,T)\frac{1-e^{-2\iota(T-t)}}{2\iota} \leq a^{(n+1)}(t) \leq \frac{1-e^{-2\iota(T-t)}}{2\iota}.$$

According to Assumption 1, this implies that $a_2^{(n+1)} \in B(M,T)$. Now let ζ be a smooth bounded function such that $\zeta(x) = x$ for $x \in [-M,T]$ and $\zeta(x) = 0$ if $x \leq -M-1$ or $x \geq T+1$. Then, it holds that

$$\dot{a}_2^* = 2\iota a_2^* + \frac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu\zeta(a_2^*(t)))^2 - 1,$$

and

$$\dot{a}_2^{(n+1)} = 2\iota a_2^{(n+1)} + \frac{\gamma^2}{(1+\gamma)^2}(1+\rho\nu\zeta(a_2^{(n)}(t)))^2 - 1.$$

We see that $a^{(n)}$ is the Picard iteration. Picard-Lindelöf theorem states that $a_2^{(n)}$ will uniformly converge to a_2^*. The convergence of $a_1^{(n)}$ immediately follows. Hence, we finish the proof. $\square$

Electronic copy available at: https://ssrn.com/abstract=3770818